Universidade de Aveiro **Departamento de Matemática**
2017

Alberto Oliveira da Silva

Biplots associados ao PLS: uma aplicação a estudos quimiométricos

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Professora Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

À minha mulher, Danielli, e aos meus filhos, Caio, Valentina e Eloá, pelo amor e compreensão, mesmo diante de minha ausência em muitos momentos importantes.

O júri

Presidente

Doutor Pedro Filipe Pessoa Macedo
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Doutora Magda Sofia Valério Monteiro
Professora Adjunta da Escola Superior de Tecnologia e Gestão de Águeda da Universidade de Aveiro

Agradecimentos

À minha orientadora, Professora Adelaide de Fátima Baptista Valente Freitas, pela sua dedicação e disponibilidade para me ensinar, mas também por ter acreditado em mim, um aluno diferente.

Ao Professor Agostinho Miguel Mendes Agra, Diretor do Mestrado em Matemática e Aplicações, pela sua ajuda em um momento delicado do meu percurso na Universidade de Aveiro, a quem serei eternamente grato.

A todos os meus Professores e Professoras do Departamento de Matemática da Universidade de Aveiro, pelos conhecimentos transmitidos.

À investigadora Iola Melissa Fernandes Duarte e à Professora Ana Maria Pissarra Coelho Gil, ambas do Departamento de Química da Universidade de Aveiro, por terem gentilmente fornecido os dados utilizados na aplicação do método estudado.

Por último, mas acima de tudo, ao Grande Arquiteto do Universo!

Palavras-chave

Mínimos quadrados parciais, PLS, *Biplot*, Modelos de regressão linear múltipla, Colinearidade.

Resumo

Estabelecer as relações lineares entre dois conjuntos de variáveis é uma tarefa que pode apresentar alguma complexidade, ainda mais se as variáveis que explicam o modelo forem fortemente correlacionadas. O desenvolvimento da tecnologia tornou este cenário ainda mais comum, uma vez que os experimentos em muitas áreas tendem a gerar cada vez mais variáveis e isso pode levar à colinearidade entre elas. Dentre os modelos que se propõem a solucionar esse tipo de problema, o Método de Mínimos Quadrados Parciais (PLS – do inglês *Partial Least Squares*) ganhou certa popularidade nos últimos anos, principalmente na Quimiometria, por ser pouco restritivo. Além disso, o PLS permite ainda a descrição da estrutura subjacente aos dados, levando a uma necessidade do desenvolvimento de técnicas exploratórias de visualização que facilitem tanto a compreensão dessa estrutura quanto do método em si. Nesse sentido, o *biplot* PLS surge para preencher essa lacuna e se tornar mais uma ferramenta à disposição dos pesquisadores. Com o propósito de dar mais consistência ao estudo, os métodos e as técnicas que fazem parte desta dissertação são postos em prática com sua aplicação a dados reais da Quimiometria, especificamente dados espectrais de ressonância magnética nuclear (RMN) de biofluidos.

Keywords Partial least squares, PLS, Biplot, Multiple linear regression model, Collinearity.

Abstract

Establishing linear relations between two sets of variables is a task that can present some complexity, especially if the variables that explain the model are strongly correlated. The development of technology has made this scenario even more common, since experiments in many areas tend to generate more and more variables and this can lead to collinearity between them. Among the models that propose to solve this type of problem, the Partial Least Squares (PLS) method has gained some popularity in recent years, mainly in Chemometrics, because it is less restrictive than others. In addition, the PLS also allows the description of underlying data structure, forcing the development of exploratory visualization techniques to make all of it easier to understand. Hence, the PLS biplot arises to fill this gap and become another tool available to researchers. Therefore, this is the object of this work and to become this study even more consistent, the methods and techniques referred in this dissertation are applied to real chemometric data, specifically nuclear magnetic resonance (NMR) spectral data of biofluids.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

O desenvolvimento tecnológico vivenciado nas últimas décadas tem permitido o surgimento de equipamentos e instrumentos capazes de gerar e armazenar grandes quantidades de dados. Nos mais diversos campos da Ciência, uma consequência conexa foi o aumento do número de variáveis associadas aos indivíduos nas amostras de experimentos, assim como do grau de complexidade na estrutura dos dados correlacionados.

Na Quimiometria, área destinada à análise de dados químicos de natureza multivariada, a magnitude da dimensão das matrizes de dados avançou rapidamente. A percepção do que era considerado um conjunto com muitas variáveis se alterou em apenas duas décadas. Do início dos anos 1980 até o início deste milênio, essa quantidade passou de algumas dezenas para a ordem de milhares de variáveis [21]. Nesse período, análises quantitativas demoradas e imprecisas, como titulação e precipitação, foram substituídas por técnicas instrumentais, como a espectroscopia e a cromatografia. Embora essas técnicas aliem velocidade e precisão, a informação obtida por elas não é fornecida diretamente, mas é alcançada a partir da quantificação de um grande número de curvas e picos, que dão origem a um igual número de variáveis.

A Biologia Molecular, por sua vez, também contribuiu para essa mudança de perspectiva na análise estatística de dados. Os resultados quantitativos obtidos com a utilização de *microarrays* de ADN, tecnologia desenvolvida nos anos 1990 no campo da Genômica Funcional, também é um exemplo de conjuntos de dados com poucas observações, mas de elevada dimensionalidade [2]. Essa grande quantidade de variáveis, associado ao fato de ser comum a ocorrência de pequenas amostras, pode levar à inviabilização da aplicação de modelos multivariados tradicionais. Esse novo cenário tem contribuído para o surgimento de outros métodos matemáticos e estatísticos, capazes de lidar com o problema apresentado e de gerar resultados mais robustos do que os métodos tradicionais [20].

1.2 Regressão linear múltipla e colinearidade

Na Análise Estatística de Dados, a regressão linear é um dos mais populares métodos utilizados. Um problema de regressão linear múltipla consiste em se obter um modelo que relacione, a menos de um erro aleatório $\boldsymbol{\varepsilon}$, uma variável dependente $\boldsymbol{Y}$, também denominada *resposta*, a um conjunto de variáveis independentes, igualmente chamadas de explicativas ou preditoras, sendo descrito por:

$$\boldsymbol{Y} = \boldsymbol{Xb} + \boldsymbol{\varepsilon},$$

onde o vetor $\boldsymbol{Y}$ tem dimensão $n\text{x}1$; a matriz $\boldsymbol{X}$ possui dimensão nxm; o vetor $\boldsymbol{b}$, dos coeficientes de regressão, tem dimensão $m\text{x}1$ (considere-se $\boldsymbol{X}$ e $\boldsymbol{Y}$ centradas) e o vetor dos resíduos $\boldsymbol{\varepsilon}$ possui dimensão $n\text{x}1$, assim como n é o número de observações e m a quantidade de variáveis independentes. No caso de haver mais do que uma variável resposta, a regressão diz-se multivariada e $\boldsymbol{Y}$ é representada por uma matriz com dimensão nxp, sendo que os coeficientes de regressão serão descritos pela matriz $\boldsymbol{B}$, de dimensão mxp, em que p é a quantidade de variáveis dependentes.

A perspectiva da existência de muitas variáveis e amostras pequenas pode, contudo, violar os pressupostos sob os quais o Modelo de Regressão Linear Múltipla (MRLM) é construído e inviabilizar sua aplicação. Por exemplo, quando se verifica uma quantidade de variáveis independentes maior do que o número de observações, ou seja, $m > n$, é fácil provar que existirá um número infinito de soluções para $\boldsymbol{b}$. Por outro lado, um elevado número m de preditores aumenta a possibilidade da ocorrência de colinearidade [1].

Quando $n > m$, uma solução para $\boldsymbol{b}$ pode ser alcançada ao se minimizar $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{Xb}$ segundo algum critério, como por exemplo, pelo método de mínimos quadrados ordinários (OLS, do inglês *Ordinary Least Squares*). Segundo o método OLS, ao se considerar o conjunto B de todos os possíveis valores de $\boldsymbol{b}$, tal que $B \subseteq \mathbb{R}^m$, pretende-se encontrar o vetor $\boldsymbol{b}' = (b_1, \dots, b_m) \in B$ que minimiza a soma dos quadrados dos resíduos dada por:

$$H(b) = \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i^2$$
$$= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$$

$$= (\boldsymbol{y} - \boldsymbol{Xb})'(\boldsymbol{y} - \boldsymbol{Xb})$$

$$= \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Xb} - 2\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y}.$$

Uma condição necessária sobre $\boldsymbol{b}$ para que $H(b)$ seja mínimo é que

$$\frac{\partial H(b)}{\partial b} = 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} - 2\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0},$$

o que resulta nas denominadas equações normais

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}.$$

Assumindo que $\boldsymbol{X}$ tenha posto completo (m), então $\boldsymbol{X}'\boldsymbol{X}$ é definida positiva e a solução do problema de regressão linear será única e obtida por

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Portanto, um dos pressupostos do MRLM é que a matriz dos preditores $\boldsymbol{X}$ tenha posto completo (m). A violação dessa condição define o problema da colinearidade [3], levando a um incremento da variância dos estimadores de mínimos quadrados [5].

Uma das maneiras de lidar com o problema da colinearidade é a construção de uma nova matriz de variáveis latentes ortogonais, também chamadas de *fatores* ou *componentes*, em que cada uma delas é uma combinação linear de $X = (X_1, \dots, X_m)$ [9]. O método de mínimos quadrados parciais, também conhecido como PLS (*Partial Least Squares*), é uma técnica estatística que permite a construção de tais variáveis, reduz a dimensão da matriz singular e contorna o contratempo da colinearidade dos preditores, uma vez que extrai fatores não correlacionados uns aos outros.

1.3 O PLS como alternativa

O PLS pode ser descrito como uma técnica que alia algumas características da Análise de Componentes Principais (ACP) com outras do MRLM [1, 17]. É considerado uma extensão do MRLM, porém mais flexível do que outros métodos que buscam contornar as suas imposições, como a regressão por componentes principais e a análise de correlação canônica, sendo o menos restritivo dos métodos multivariados. Portanto, o PLS é aplicável mesmo que os dados possuam uma grande quantidade de variáveis fortemente correlacionadas, sendo capaz de estimar as relações lineares entre as variáveis X e Y e fazer predições para novos dados observados.

Ao longo deste trabalho serão considerados dois conjuntos de variáveis, $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_p)$ e $\boldsymbol{X} = (X_1, X_2, \dots, X_m)$, onde para cada uma das variáveis existem n observações, assim como $m > n$.

As seguintes matrizes são estabelecidas no desenvolvimento do método PLS:

$\boldsymbol{Y}_o$: matriz de dados originais das variáveis resposta, com dimensão nxp

$\boldsymbol{X}_o$: matriz de dados originais das variáveis explicativas, com dimensão nxm

As matrizes de dados originais serão centradas na média para uniformizar os dados e facilitar os cálculos, de tal forma que:

$$\boldsymbol{Y} = \boldsymbol{Y}_o - \mathbf{1}\boldsymbol{z}'\boldsymbol{Y}_o = (\boldsymbol{I} - \mathbf{1}\boldsymbol{z}')\boldsymbol{Y}_o$$

$$\boldsymbol{X} = \boldsymbol{X}_o - \mathbf{1}\boldsymbol{z}'\boldsymbol{X}_o = (\boldsymbol{I} - \mathbf{1}\boldsymbol{z}')\boldsymbol{X}_o$$

onde $\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{nx1}$, $\boldsymbol{z}' = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})_{1xn}$ e $\boldsymbol{I}$ é a matriz identidade com dimensão nxn.

Será demonstrado que o PLS busca estimar a relação linear entre $\boldsymbol{Y}$ e $\boldsymbol{X}$ através de um novo conjunto de variáveis latentes $\boldsymbol{T} = (T_1, T_2, \dots, T_k)$, que são escritas como combinações lineares dos preditores. Além disso, a correlação entre esses fatores é nula, ou seja, $\mathrm{Cor}(T_i, T_j) = 0, \forall\, i \neq j$ e o posto de $\boldsymbol{T}$ é menor do que o posto de $\boldsymbol{X}$. As variáveis latentes $T_1, T_2, \dots, T_k$, também denominadas como componentes PLS, são obtidas pela decomposição simultânea de $\boldsymbol{Y}$ e $\boldsymbol{X}$, e tem como resultado as seguintes relações:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}' + \boldsymbol{F}.$$

As matrizes $\boldsymbol{T}_{\mathrm{nxk}}$, $\boldsymbol{P}_{\mathrm{mxk}}$, $\boldsymbol{E}_{\mathrm{nxm}}$ representam, respectivamente, os *scores*[1], *loadings* e resíduos associados aos preditores $\boldsymbol{X}$ e $\boldsymbol{U}_{\mathrm{nxk}}$, $\boldsymbol{Q}_{\mathrm{pxk}}$, $\boldsymbol{F}_{nxp}$ são as matrizes de *scores*, *loadings* e resíduos relativos à resposta $\boldsymbol{Y}$. Como também será mostrado, a escolha da quantidade k de componentes a serem extraídas pode ser definida pela técnica denominada validação cruzada.

Para a aplicação do modelo, presume-se que $\boldsymbol{X}$ terá em suas colunas m preditores (fortemente) correlacionados, sendo $m > n$. Em relação à matriz $\boldsymbol{Y}$, se $p = 1$, o modelo é

[1] Define-se como *score* o valor da projeção ortogonal de uma observação da amostra sobre a direção de uma componente e o *loading* como o cosseno do ângulo formado entre uma componente e o eixo de uma variável específica.

chamado de PLS1 (univariado) e a extração dos fatores ocorre somente em $\boldsymbol{X}$. Caso $p \geq 2$, para o qual a literatura adotou o nome PLS2, a extração das componentes é feita por intermédio de algoritmos iterativos de forma simultânea em ambas as matrizes de dados $\boldsymbol{X}$ e $\boldsymbol{Y}$, com o propósito de que esses vetores latentes expliquem tanto o quanto possível a covariância entre os dois blocos. Apesar de existirem diversos algoritmos que calculem as variáveis latentes no PLS, apenas o NIPALS (acrônimo do inglês *Nonlinear Iterative Partial Least Squares*) é abordado neste trabalho por ser o algoritmo original do método.

1.4 A visualização de técnicas multivariadas e do PLS

Na análise multivariada de elevada dimensionalidade, a compreensão teórica das relações existentes entre as variáveis e de sua estrutura subjacente pode ser suplementada pela visualização dos dados e dos resultados da aplicação do método utilizado. A exibição gráfica tem a capacidade de sumariar todo o conjunto de maneira simplificada, sendo uma importante ferramenta de natureza exploratória, pois intensifica a percepção do investigador de um modo mais subjetivo do que quantitativo. Esse apelo visual torna mais fácil a assimilação daquilo que é representado, funcionando como complemento à frieza do número puramente [14].

A visualização gráfica de um conjunto de dados multivariados por intermédio de seu *biplot* cumpre essa função. Gabriel [8] propôs que qualquer matriz de dados pode ter seus elementos calculados como produto escalar de dois vetores que representam as linhas e as colunas de duas outras matrizes. Se a matriz de dados tem posto 2, ou pode ser aproximada por uma matriz com tal posto, então esses vetores podem ser representados graficamente. No PLS *biplot*, a matriz $\boldsymbol{X}$ é aproximada, a menos de um termo residual $\boldsymbol{E}$, por

$$\widetilde{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{P}'.$$

Considerando que cada elemento de $\widetilde{\boldsymbol{X}}$ pode ser expresso pelo produto escalar das linhas de $\boldsymbol{T}$ e $\boldsymbol{P}$, o gráfico PLS *biplot* é então traçado com base nessa propriedade. Da mesma maneira, essa abordagem pode ser utilizada tanto em relação à matriz de aproximação de $\boldsymbol{Y}$ quanto à matriz (ou vetor) dos coeficientes de regressão para, ao final, se obter um gráfico das relações e estruturas existentes no modelo.

1.5 Objetivos da dissertação

O objetivo geral da presente dissertação é apresentar a formulação teórica da técnica de visualização *biplot* do método PLS, na esteira do que foi proposto no trabalho elaborado por [17]. Está contida nesse propósito, a demonstração da operacionalidade do método PLS por intermédio do algoritmo NIPALS, assim como a construção de técnicas exploratórias do *biplot*, como a calibragem e delimitação de áreas nos gráficos. Como objetivo específico, pretende-se mostrar a sua funcionalidade com uma aplicação à Quimiometria em dados reais.

1.6 Organização da dissertação

Este capítulo começa por mostrar as circunstâncias fáticas e técnicas que, no contexto da regressão linear multivariada, motivaram o surgimento de métodos capazes de lidar com colinearidade entre variáveis preditoras. Foi feita uma apresentação inicial da regressão por mínimos quadrados parciais, que é um método específico para tratar o problema colocado, assim como de seu *biplot* como técnica exploratória dos resultados.

O segundo capítulo é dedicado a fornecer as explicações necessárias para a compreensão do PLS, nomeadamente os conceitos básicos e o desenvolvimento dos cálculos envolvidos na especificação das variáveis latentes, pesos, *loadings* e *scores*. Além disso, discute-se a validação do modelo feita por intermédio da validação cruzada, sendo que para sua demonstração é utilizado o conjunto de dados do pacote *pls* do R denominado *gasoline* [4].

No terceiro capítulo são dadas as noções básicas para o entendimento da técnica de visualização *biplot* para, após, aplicá-la ao PLS. É também mostrada a técnica de calibragem dos eixos *biplot*, que permite se obter os dados das variáveis aproximados pelo PLS e os coeficientes de regressão do modelo diretamente no gráfico. Para completar, é feita a apresentação do método das áreas *biplot*, utilizado para se estimar visualmente os coeficientes de regressão PLS. A exemplificação no segundo capítulo utiliza o conjunto de dados do pacote *pls* do R denominado *oliveoil*.

O quarto capítulo é reservado para a aplicação dos métodos estudados a dados reais de natureza quimiométrica. Essa base de dados contém dados espectrais de ressonância magnética nuclear (RMN) relativos a fluidos humanos, denominada neste trabalho como *base Plasma*. A parte final da dissertação está no quinto capítulo, onde constam as conclusões.

Toda a parte de programação computacional envolvida foi desenvolvida por meio de implementação dos algoritmos no ambiente R [18] e os códigos utilizados se encontram descritos no Apêndice.

Capítulo 2

O Método PLS

2.1 Contexto histórico e definições

O método PLS, acrônimo de *partial least squares*, tem origem em estudos de Econometria com o trabalho elaborado por [22], mas ganhou popularidade com trabalhos publicados na área da Quimiometria duas décadas mais tarde, como [12, 16]. O PLS é uma técnica da análise multivariada de dados que combina outros dois métodos, nomeadamente, a análise de componentes principais (ACP) e a regressão linear múltipla. Tem como objetivo predizer variáveis respostas $\boldsymbol{Y}$ (um bloco) através de um modelo baseado em combinações lineares de variáveis explanatórias $\boldsymbol{X}$ (outro bloco), levando-se em consideração a estrutura comum entre os dois blocos de variáveis.

Quando $\boldsymbol{Y}$ é univariado e $\boldsymbol{X}$ tem posto completo, o método de mínimos quadrados ordinários (OLS, acrônimo de *ordinary least squares*) é capaz de fornecer uma solução para o problema de regressão linear multivariada. Contudo, quando o número de variáveis explicativas é elevado e maior do que a quantidade de observações, é provável que $\boldsymbol{X}$ seja quase singular devido à colinearidade, fazendo com que a solução obtida pelo OLS tenha grande variabilidade e se torne instável [15], definindo o problema como mal posto. O PLS supera o problema da forte correlação entre as variáveis explicativas ao extrair de ambos os blocos, simultaneamente (Figura 2.1), novos conjuntos de variáveis latentes ortogonais. Estas novas variáveis são capazes de explicar a estrutura de variância de cada bloco e também a covariância entre eles, tanto quanto possível. Como já mencionado na Seção 1.3 do Capítulo 1, as matrizes de aproximação são dadas por:

$$\boldsymbol{X} \approx \boldsymbol{T}\boldsymbol{P}'$$

$$\boldsymbol{Y} \approx \boldsymbol{U}\boldsymbol{Q}'$$

sendo que os *scores* de $\boldsymbol{X}$ e os *scores* de $\boldsymbol{Y}$, concretamente $\boldsymbol{T}$ e $\boldsymbol{U}$, relacionam-se internamente (Figura2.1), segundo um modelo de regressão linear, da forma:

$$\boldsymbol{U} = \boldsymbol{T}\boldsymbol{B} + \boldsymbol{H}.$$

Aqui, $\boldsymbol{B}$ é uma matriz diagonal de ordem k (quantidade de componentes principais retidas no modelo), cujos elementos não-nulos são os coeficientes de regressão b_i da relação linear interna e $\boldsymbol{H}$ a matriz dos resíduos. A matriz $\boldsymbol{B}$ estimada pelo método OLS fornece uma estimativa para $\boldsymbol{U}$:

$$\widehat{\boldsymbol{U}} = \boldsymbol{T}\widehat{\boldsymbol{B}}.$$

Para se garantir a máxima covariância entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ ao se extrair as componentes PLS, faz-se necessário encontrar dois conjuntos de pesos $\boldsymbol{w}$ e $\boldsymbol{q}$, de tal forma que os seguintes vetores sejam obtidos

$$\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$$

$$\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{q}.$$

Uma forma de se obter $\mathrm{COV}(\mathrm{X}, \mathrm{Y})$ máxima é fazer com que $\boldsymbol{t}'\boldsymbol{u}$ seja máximo, o que pode se resumir a um problema de otimização, do tipo

$$\underset{\boldsymbol{w},\boldsymbol{q}}{\operatorname{argmax}} \{\boldsymbol{w}'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{q}\},$$

sujeito a restrições em relação aos vetores $\boldsymbol{w}$ e $\boldsymbol{t}$, tal que: COR $(t_i, t_j) = 0$ e COR $(w_i, w_j) = 0$, $\forall\, i \neq j$, e ainda que $\boldsymbol{t}'\boldsymbol{t} = 1$ e $\boldsymbol{w}'\boldsymbol{w} = 1$.

Como será mostrado em seguida, uma das maneiras para se obter uma solução para esse problema é através do algoritmo NIPALS (acrônimo do inglês *Nonlinear Iterative Partial Least Squares*).

2.2 O algoritmo NIPALS do caso univariado

2.2.1 Aspectos gerais

No PLS1 existe apenas uma variável resposta e o objetivo do método é predizer $\boldsymbol{y}$ por meio de transformações lineares das variáveis explicativas, o que resulta no seguinte modelo de regressão linear:

$$\hat{y} = \hat{b}_1 T_1 + \hat{b}_2 T_2 + \cdots + \hat{b}_k T_k \,,$$

onde T_k , $k = 1, \ldots, K$, são combinações lineares dos preditores $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$, e $COR(T_i, T_j) = 0, \forall i \neq j$. Tratando-se do caso univariado e considerando n observações, a matriz da variável resposta terá dimensão nx1 e a matriz dos preditores terá dimensão nxm. A decomposição desta última resulta em $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E}$.

Neste caso, $\boldsymbol{T} = (\boldsymbol{t}_1 \ \boldsymbol{t}_2 \ \dots \ \boldsymbol{t}_k)$, em que $\boldsymbol{t}_i$, $i = 1, \dots, k$, é um vetor coluna com dimensão nx1; $\boldsymbol{P} = (\boldsymbol{p}_1 \ \boldsymbol{p}_2 \ \dots \ \boldsymbol{p}_k)$, em que $\boldsymbol{p}_i$, $i = 1, \dots, k$, é um vetor coluna com dimensão mx1; e $\boldsymbol{E}$ é a matriz dos resíduos, cuja dimensão é nxm. Em sua primeira iteração, o NIPALS calcula os vetores $\boldsymbol{t}_1$ e $\boldsymbol{p}_1$ e, em seguida, a primeira matriz de resíduos $\boldsymbol{E}_1 = \boldsymbol{X} - \boldsymbol{t}_1\boldsymbol{p}_1'$. Na iteração seguinte, o resíduo $\boldsymbol{E}_1$ é utilizado para calcular $\boldsymbol{t}_2$ e $\boldsymbol{p}_2$ para, com estes, obter-se o resíduo $\boldsymbol{E}_2 = \boldsymbol{E}_1 - \boldsymbol{t}_2\boldsymbol{p}_2'$. Assim, a forma recursiva é dada por [10]:

$$\boldsymbol{E}_h = \boldsymbol{E}_{h-1} - \boldsymbol{t}_h\boldsymbol{p}_h' \ ,$$

$h = 1, \dots, r$, onde $r = posto(\boldsymbol{X})$ e $\boldsymbol{X} = \boldsymbol{E}_0$.

O algoritmo NIPALS possui uma fase de pré-processamento, na qual as variáveis são centradas[1] com base na média das colunas e é então atribuída uma nova designação [9]:

$$\boldsymbol{F}_0 = \boldsymbol{Y}$$

$$\boldsymbol{E}_{0j} = \boldsymbol{X}_j, \ j = 1, \dots, m.$$

Referenciando-se aos dados, obtém-se os vetores $\boldsymbol{f}_0$ e $\boldsymbol{e}_{oj}$, ambos com dimensão nx1, tal que:

$$\boldsymbol{f}_0 = \boldsymbol{y} \ , \ (\bar{f}_0 = \bar{y} = 0)$$

$$\boldsymbol{e}_{0j} = \boldsymbol{x}_j \ , \ (\bar{e}_{0j} = \bar{x}_j = 0).$$

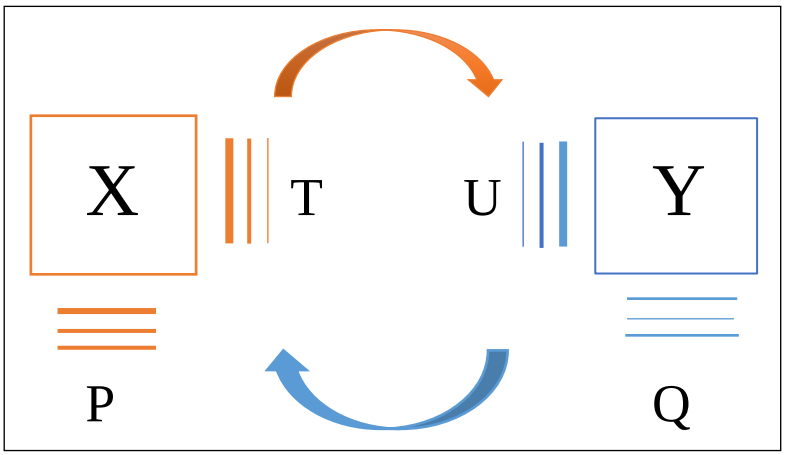

Figura 2.1: Diagrama representativo do PLS, conforme [21].

[1] Note-se que, conforme a Seção 1.3 do Capítulo 1, *X* e *Y* designam as variáveis originais (X_o e Y_o) centradas.

2.2.2 Algoritmo NIPALS univariado (PLS1)

Passo 1. Calcular as matrizes centradas $\boldsymbol{X}$ e $\boldsymbol{Y}$ a partir das matrizes originais $\boldsymbol{X_o}$ e $\boldsymbol{Y_o}$

Fazer $\boldsymbol{E}_0 = \boldsymbol{X}$ e $\boldsymbol{F}_0 = \boldsymbol{Y}$

Inicializar $\boldsymbol{T} = [\,]$, $\boldsymbol{P} = [\,]$, $\boldsymbol{W} = [\,]$ e $\boldsymbol{B} = [\,]$.

Passo 2. Para $k = 1$ até K ($K \leq posto(\boldsymbol{X})$), repetir:

2.1 $\boldsymbol{w}_k^* = \boldsymbol{E}'_{k-1}\boldsymbol{f}_{k-1}$ (Determinação do vetor de pesos)

2.2 $\boldsymbol{w}_k = \frac{\boldsymbol{w}_k^*}{\|\boldsymbol{w}_k^*\|}$ (Normalização do vetor de pesos)

2.3 $\boldsymbol{t}_k^* = \boldsymbol{E}_{k-1}\boldsymbol{w}_k$ (Cálculo do vetor de *scores*)

2.4 $\boldsymbol{t}_k = \frac{\boldsymbol{t}_k^*}{\|\boldsymbol{t}_k^*\|}$ (Normalização do vetor de *scores*)

2.5 $b_k = \frac{\boldsymbol{t}_k'\boldsymbol{f}_{k-1}}{\boldsymbol{t}_k'\boldsymbol{t}_k} = \boldsymbol{t}_k'\boldsymbol{f}_{k-1}$ (Coeficiente de regressão da relação interna)

2.6 $\boldsymbol{p}_k = \frac{\boldsymbol{E}'_{k-1}\boldsymbol{t}_k}{\boldsymbol{t}_k'\boldsymbol{t}_k} = \boldsymbol{E}'_{k-1}\boldsymbol{t}_k$ (Cálculo do vetor de *loadings*)

2.7 $\boldsymbol{E}_k = \boldsymbol{E}_{k-1} - \boldsymbol{t}_k\boldsymbol{p}'_k$ (Deflação da matriz dos preditores)

2.8 $\boldsymbol{F}_k = \boldsymbol{F}_{k-1} - \boldsymbol{t}_k b_k$ (Deflação do vetor da resposta)

2.9. Atualizar as matrizes e vetores:

$\boldsymbol{T} = [\boldsymbol{t}_k]$

$\boldsymbol{W} = [\boldsymbol{w}_k]$

$\boldsymbol{P} = [\boldsymbol{p}_k]$

$\boldsymbol{b} = (b_k)$.

2.2.3 Explicações sobre o algoritmo NIPALS PLS1

O algoritmo NIPALS pode ser explicado sob o ponto de vista matemático por meio da projeção de vetores em cada um de seus passos. Neste estudo, contudo, será adotada a abordagem estatística, por intermédio da qual a regressão linear é a ferramenta utilizada para se construir o modelo. O passo-a-passo do algoritmo pode ser descrito como abaixo:

1. Os passos *2.1* e *2.3* do algoritmo NIPALS PLS1 baseiam-se na realização da regressão de $\boldsymbol{F}_0$ sobre cada uma das variáveis $\boldsymbol{E}_{0j}$, separadamente, obtendo-se m equações de regressão univariadas, tal que:

$$\widehat{\boldsymbol{F}}_0 = a_{0j}\boldsymbol{E}_{0j}\ ,\ j = 1, \dots, m$$

$$\text{com } a_{0j} \text{ estimado por } \frac{\boldsymbol{e}'_{0j}\boldsymbol{f}_0}{\boldsymbol{e}'_{0j}\boldsymbol{e}_{0j}}.$$

A primeira componente T_1 é obtida por intermédio de uma média ponderada dessas m equações, utilizando-se os pesos $z_{0j} = \boldsymbol{e}'_{0j}\boldsymbol{e}_{0j}$, $j = 1, \dots, m$, que são proporcionais à variância de cada variável centrada, resultando em:

$$\boldsymbol{t}_1 = \sum_{j=1}^{m} z_{0j}\, a_{0j}\boldsymbol{E}_{0j} =$$

$$\sum_{j=1}^{m} \boldsymbol{e}'_{0j}\boldsymbol{e}_{0j}\frac{\boldsymbol{e}'_{0j}\boldsymbol{f}_0}{\boldsymbol{e}'_{0j}\boldsymbol{e}_{0j}}\boldsymbol{E}_{0j} =$$

$$\sum_{j=1}^{m} \boldsymbol{e}'_{0j}\boldsymbol{f}_0\ \boldsymbol{E}_{0j}.$$

Fazendo $w_{1j} = \boldsymbol{e}'_{0j}\boldsymbol{f}_0$ (Passo *2.1*), resulta em (Passo *2.3*):

$$\boldsymbol{t}_1 = \sum_{j=1}^{m} w_{1j}\,\boldsymbol{E}_{0j}.$$

Com o propósito de simplificar a demonstração, considera-se que o vetor $\boldsymbol{t}_1$ deva ser normalizado nesse momento (Passo *2.4*), embora não seja necessário segundo algumas versões do NIPALS. Assim, a nova variável latente T_1 é um preditor para F_0 e a relação linear interna do modelo é estabelecida por (Passo *2.5*):

$$F_0 = b_1 T_1 + \boldsymbol{\varepsilon}_1$$

$$\text{com } b_1 \text{estimado por } \frac{\boldsymbol{t}_1' \boldsymbol{f}_0}{\boldsymbol{t}_1' \boldsymbol{t}_1} = \boldsymbol{t}_1' \boldsymbol{f}_0.$$

2. Existe, todavia, informação contida nas variáveis preditoras que não foram captadas pela primeira componente $\boldsymbol{T}_1$. Essa quantidade pode ser estimada por meio do resíduo da regressão de $\boldsymbol{E}_{0j}$ sobre $\boldsymbol{T}_1$, que será nomeada como $\boldsymbol{E}_{1j}$, como abaixo (Passos *2.6* e *2.7*):

$$\boldsymbol{E}_{1j} = \boldsymbol{E}_{0j} - p_{1j} \boldsymbol{T}_1$$

$$\text{com } p_{1j} \text{ estimado por } \frac{\boldsymbol{t}_1' \boldsymbol{e}_{0j}}{\boldsymbol{t}_1' \boldsymbol{t}_1} = \boldsymbol{t}_1' \boldsymbol{e}_{0j}.$$

É nessa etapa que ocorre a deflação da matriz das variáveis explicativas, tendo como resultado a primeira matriz de resíduos. Do mesmo modo, para se estimar a quantidade de variabilidade da variável resposta que não foi explicada por T_1, procede-se à deflação de $\boldsymbol{F}_0$ e o resíduo passa a ser chamado de $\boldsymbol{F}_1$(Passo *2.8*):

$$\boldsymbol{F}_1 = \boldsymbol{F}_0 - b_1 \boldsymbol{T}_1.$$

3. Procede-se similarmente a regressão de F_1 sobre cada uma das variáveis deflacionadas E_{1j}, tal que:

$$\hat{F}_1 = a_{1j} E_{1j} \ , \ j = 1, \dots, m$$

$$\text{com } a_{1j} \text{ estimado por } = \frac{\boldsymbol{e}_{1j}' \boldsymbol{f}_1}{\boldsymbol{e}_{1j}' \boldsymbol{e}_{1j}}.$$

Mais uma vez, a segunda componente T_2 é obtida pela média ponderada das m equações de regressão, utilizando-se a variância de cada preditor como peso inicial, mas que, após simplificações, resultará no peso $w_{2j} = \boldsymbol{e}_{1j}' \boldsymbol{f}_1$, fazendo com que:

$$\boldsymbol{t}_2 = \sum_{j=1}^{m} w_{2j} \, \boldsymbol{E}_{1j}.$$

4. Considerando que $\boldsymbol{X}$ tem posto r, o algoritmo pode ser repetido até r iterações, altura em que a matriz deflacionada $\boldsymbol{X}$ torna-se nula. Entretanto, o número de

componentes extraídos (k) é, em regra, inferior ao número máximo possível de iterações, de tal modo que o modelo pode ser escrito como

$$F_0 = b_1T_1 + b_2T_2 + \cdots + b_kT_k + \varepsilon_k \text{ , sendo } F_0 = Y.$$

Os vetores dos *loadings* ($\boldsymbol{p}_i$) relativos às variáveis explicativas são, portanto, coeficientes da regressão realizada para estimar a quantidade de informação de cada uma das variáveis que está presente em determinada componente. O *loading* fornece, em certa medida, a importância do preditor para o modelo em termos de sua variância [6]. A utilização dos vetores $\boldsymbol{p}_i$ para deflacionar a matriz dos preditores é que assegura a ortogonalidade dos $\boldsymbol{T}_i$ [6, 10].

Os coeficientes de regressão b_i estabelecem a relação linear interna do modelo e fornecem a importância de cada componente para explicar a variabilidade da variável resposta. Eles também são utilizados na determinação dos coeficientes de regressão do PLS ($\widehat{\boldsymbol{b}}_{PLS}$) para a predição da variável dependente Y. Os vetores de pesos $\boldsymbol{w}$ constituem a direção de maior variabilidade de $\boldsymbol{X}'\boldsymbol{y}$ e a projeção das observações sobre esse eixo determina os scores $\boldsymbol{t}$ [6].

2.3 O algoritmo NIPALS do caso multivariado

No PLS2, existem duas ou mais variáveis respostas, ou seja, a matriz $\boldsymbol{Y}$ tem dimensão nxp, em que $p \geq 2$. O algoritmo funciona da mesma forma que no caso univariado, apenas com algumas adaptações, dado que neste caso são extraídas, simultaneamente, componentes tanto em $\boldsymbol{X}$ quanto em $\boldsymbol{Y}$, nomeadamente, $\boldsymbol{t}$ e $\boldsymbol{u}$.

Como no PLS1, as matrizes de dados originais são centradas pela média das colunas[2] e renomeadas:

$$\boldsymbol{F}_{0i} = \boldsymbol{Y}_i \text{ , } i = 1, \dots, p$$

$$\boldsymbol{E}_{0j} = \boldsymbol{X}_j \text{ , } j = 1, \dots, m.$$

[2] Relembrando que, conforme Seção 1.3, $\boldsymbol{X}$ e $\boldsymbol{Y}$ representam as matrizes de dados das variáveis já centradas, dado que: $\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{z}')\boldsymbol{X}_o$ e $\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{z}')\boldsymbol{Y}_o$, sendo $\boldsymbol{X}_o$ e $\boldsymbol{Y}_o$ os dados originais. Assim, $\boldsymbol{X}_j$ e $\boldsymbol{Y}_i$ são colunas centradas de $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente.

2.3.1 Algoritmo NIPALS PLS2

Passo 1. Centrar $\boldsymbol{X_o}$ e $\boldsymbol{Y_o}$;
Fazer $\boldsymbol{E}_0 = \boldsymbol{X}$, $\boldsymbol{F}_0 = \boldsymbol{Y}$ e $\boldsymbol{u}_1 = \boldsymbol{f}_{0i}$ (coluna i de $\boldsymbol{F}_0$ com maior variância)
Inicializar $\boldsymbol{T} = [\,]$, $\boldsymbol{P} = [\,]$, $\boldsymbol{W} = [\,]$, $\boldsymbol{U} = [\,]$, $\boldsymbol{Q} = [\,]$ e $\boldsymbol{B} = [\,]$

Passo 2. Para $k = 1$ até K ($K \leq posto(\boldsymbol{X})$), repetir:

2.1. Enquanto $\boldsymbol{t}_k$ não convergir, repetir:

2.1.1 $\boldsymbol{w}_k^* = \boldsymbol{E}_{k-1}'\boldsymbol{u}_k$ (Determinação do vetor de pesos – bloco $\boldsymbol{X}$)

2.1.2 $\boldsymbol{w}_k = \frac{\boldsymbol{w}_k^*}{\|\boldsymbol{w}_k^*\|}$ (Normalização do vetor de pesos – bloco $\boldsymbol{X}$)

2.1.3 $\boldsymbol{t}_k^* = \boldsymbol{E}_{k-1}\boldsymbol{w}_k$ (Cálculo do vetor de *scores* candidato - bloco $\boldsymbol{X}$)

2.1.4 $\boldsymbol{t}_k = \frac{\boldsymbol{t}_k^*}{\|\boldsymbol{t}_k^*\|}$ (Normalização do vetor de *scores* candidato)

2.1.5 $\boldsymbol{q}_k = \boldsymbol{F}_{k-1}'\boldsymbol{t}_k$ (Cálculo dos pesos do bloco $\boldsymbol{Y}$)

2.1.6 $\boldsymbol{u}_k = \boldsymbol{F}_{k-1}\boldsymbol{q}_k$ (Vetor de *scores* candidato - bloco $\boldsymbol{Y}$)

2.1.7 $\|\boldsymbol{t}_{old} - \boldsymbol{t}_{new}\| / \|\boldsymbol{t}_{new}\| < \varepsilon$ (critério de paragem)

2.1.8 $\boldsymbol{u}_{k+1} = \boldsymbol{u}_k$

2.2. Deflacionar as matrizes

2.2.1 $\boldsymbol{p}_k = \frac{\boldsymbol{E}_{k-1}'\boldsymbol{t}_k}{\boldsymbol{t}_k'\boldsymbol{t}_k} = \boldsymbol{E}_{k-1}'\boldsymbol{t}_k$ (Cálculo do vetor de loadings)

2.2.2 $\boldsymbol{E}_k = \boldsymbol{E}_{k-1} - \boldsymbol{t}_k\boldsymbol{p}_k'$ (Deflação da matriz do preditores)

2.2.3 $b_k = \frac{\boldsymbol{t}_k'\boldsymbol{u}_k}{\boldsymbol{t}_k'\boldsymbol{t}_k} = \boldsymbol{t}_k'\boldsymbol{u}_k$ (Coeficientes da relação interna)

2.2.4 $\boldsymbol{F}_k = \boldsymbol{F}_{k-1} - \boldsymbol{b}_k\,\boldsymbol{t}_k\boldsymbol{q}_k'$ (Deflação da matriz das respostas)

2.3. Atualizar $\boldsymbol{T}$, $\boldsymbol{P}$, $\boldsymbol{W}$, $\boldsymbol{U}$, $\boldsymbol{Q}$ e $\boldsymbol{B}$.

2.3.2 Explicações sobre o algoritmo NIPALS PLS2

Tal como no caso univariado, a seguir está descrito o passo-a-passo do algoritmo NIPALS quando existe mais de uma variável dependente.

1. Determinação das componentes e dos pesos (Passo *2.1*):

   Primeiramente, escolhe-se uma das variáveis dependentes como *proxy* da primeira componente u_1, preferencialmente a coluna da matriz Y com maior variabilidade e que será o ponto de partida nesse problema de otimização. Uma vez que a escolha tenha recaído sobre um determinado $\boldsymbol{f}_{0i}$, faz-se a sua atribuição à primeira componente candidata[3], tal que $\boldsymbol{u}_1^* = \boldsymbol{f}_{0i}$. Em seguida, é realizada a regressão linear de u_1^* sobre cada uma das variáveis explicativas centradas E_{0j}, separadamente. Tal qual no PLS1, o resultado é:

$$\hat{u}_1^* = a_{0j} E_{0j} \ , \ j = 1, \dots, m$$

$$\text{com } a_{0j} \text{ estimado por } \frac{\boldsymbol{e}_{0j}' \boldsymbol{u}_1^*}{\boldsymbol{e}_{0j}' \boldsymbol{e}_{oj}}.$$

   Em seguida, é calculada a média ponderada das m equações de regressão, atribuindo-se como peso uma medida proporcional à variância de cada uma das variáveis explanatórias, tal que $z_{0j} = \boldsymbol{e}_{0j}' \boldsymbol{e}_{oj}$. Essa média ponderada corresponde à componente candidata, que é estimada por:

$$\boldsymbol{t}_1^* = \sum_{j=1}^{m} z_{0j} \, a_{0j} \boldsymbol{E}_{0j} =$$

$$\sum_{j=1}^{m} \boldsymbol{e}_{0j}' \boldsymbol{e}_{oj} \frac{\boldsymbol{e}_{0j}' \boldsymbol{u}_1^*}{\boldsymbol{e}_{0j}' \boldsymbol{e}_{oj}} \boldsymbol{E}_{0j} =$$

$$\sum_{j=1}^{m} \boldsymbol{e}_{0j}' \boldsymbol{u}_1^* \boldsymbol{E}_{0j}.$$

[3] Aqui, a notação "*" é utilizada para designar as variáveis candidatas, ou seja, antes da convergência.

Designando o vetor $\boldsymbol{w}$ candidato como $w_{1j}^{*} = \boldsymbol{e}_{0j}'\boldsymbol{u}_1^{*}$ (Passo *2.1.1*), esses serão os coeficientes da combinação linear das variáveis explicativas que definem a componente candidata, no caso (Passo *2.1.3*):

$$\boldsymbol{t}_1^{*} = \sum_{j=1}^{m} w_{1j}^{*}\, \boldsymbol{E}_{0j}.$$

O segundo passo é fazer o caminho inverso para renovar $\boldsymbol{u}_1^{*}$, a primeira componente candidata de $\boldsymbol{Y}$. Para isso, é realizada a regressão linear de $\boldsymbol{t}_1^{*}$ sobre cada uma das variáveis dependentes centradas, separadamente, resultando em p equações de regressão:

$$\hat{t}_1 = c_{0i}F_{0i}\ ,\ \ i = 1, \dots, p$$

$$\text{com } c_{0i} \text{ estimado por } \frac{\boldsymbol{f}_{0i}'\boldsymbol{t}_1^{*}}{\boldsymbol{f}_{0i}'\boldsymbol{f}_{0i}}.$$

A partir desse ponto, é calculada a média ponderada das p equações, utilizando-se como pesos os valores $g_{0i} = \boldsymbol{f}_{0i}'\boldsymbol{f}_{0i}$, que são proporcionais à variância de cada variável dependente, e o resultado fornece uma melhor estimativa para a componente candidata $\boldsymbol{u}_1^{*}$:

$$\boldsymbol{u}_1^{*} = \sum_{i=1}^{p} g_{0i}\, c_{0i}\boldsymbol{F}_{0i} =$$

$$\sum_{i=1}^{p} \boldsymbol{f}_{0i}'\boldsymbol{f}_{0i}\frac{\boldsymbol{f}_{0i}'\boldsymbol{t}_1^{*}}{\boldsymbol{f}_{0i}'\boldsymbol{f}_{0i}}\boldsymbol{F}_{0i} =$$

$$\sum_{i=1}^{p} \boldsymbol{f}_{0i}'\boldsymbol{t}_1^{*}\ \boldsymbol{F}_{0i}.$$

Os coeficientes da combinação linear das variáveis dependentes que determinam a componente $\boldsymbol{u}_1^{*}$ são dados por $q_{0i}^{*} = \boldsymbol{f}_{0i}'\boldsymbol{t}_1^{*}$ (Passo *2.1.5*) e, portanto (Passo *2.1.6*):

$$\boldsymbol{u}_1^{*} = \sum_{i=1}^{p} q_{0i}^{*}\, \boldsymbol{F}_{0i}.$$

Retorna-se ao Passo *2.1.1* e, com a nova componente $\boldsymbol{u}_1^*$, renova-se também a componente candidata $\boldsymbol{t}_1^*$. Esse processo de melhoria das componentes se repete até que se verifique a convergência, ou em outras palavras, que o ganho seja negligenciável. A convergência é testada ao fim de cada iteração por (Passo *2.1.7*):

$$\left\|\boldsymbol{t}_{1(anterior)}^* - \boldsymbol{t}_{1(novo)}^*\right\| / \left\|\boldsymbol{t}_{1(novo)}^*\right\| < \varepsilon.$$

Após $\boldsymbol{t}_1^*$ ter convergido, os então vetores candidatos $\boldsymbol{t}_1^*$, $\boldsymbol{u}_1^*$, $\boldsymbol{w}_1^*$ e $\boldsymbol{q}_1^*$ mais recentes passam a ser designados como $\boldsymbol{t}_1$, $\boldsymbol{u}_1$, $\boldsymbol{w}_1$ e $\boldsymbol{q}_1$ e são guardados nas matrizes correspondentes $\boldsymbol{T}$, $\boldsymbol{U}$, $\boldsymbol{W}$ e $\boldsymbol{Q}$.

2. Deflação das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ (Passo *2.2*):

A regressão das variáveis explicativas centradas sobre T_1 fornece a quantidade de informação captada pela componente PLS, sendo que o vetor de *loadings* $\boldsymbol{p}$ é fornecido pelos coeficientes da regressão e a matriz dos preditores deflacionada é dada pelos resíduos. Portanto, dado que $\boldsymbol{E}_0$ é exatamente a matriz $\boldsymbol{X}$ centrada, a primeira deflação irá produzir $\boldsymbol{E}_1$, que terá como colunas (Passos *2.2.1* e *2.2.2*):

$$\boldsymbol{E}_{1j} = \boldsymbol{E}_{0j} - p_{1j}\boldsymbol{T}_1$$

$$\text{com } p_{1j} \text{ estimado por } \frac{\boldsymbol{t}_1'\boldsymbol{e}_{0j}}{\boldsymbol{t}_1'\boldsymbol{t}_1} = \boldsymbol{t}_1'\boldsymbol{e}_{0j}.$$

Com o vetor dos *scores* determinados no Passo *2.1*, a relação linear interna do modelo é estabelecida pela regressão de U_1 sobre T_1, tal que (Passo *2.2.3*):

$$U_1 = b_1 T_1 + \varepsilon_1$$

$$\text{com } b_1 \text{ estimado por } \frac{\boldsymbol{t}_1'\boldsymbol{u}_1}{\boldsymbol{t}_1'\boldsymbol{t}_1} = \boldsymbol{t}_1'\boldsymbol{u}_1.$$

Ao contrário do que é feito na deflação da matriz dos preditores, quando a matriz $\boldsymbol{t}_1\boldsymbol{p}_1'$ é subtraída de $\boldsymbol{E}_0$, para se deflacionar a matriz das variáveis dependentes não é utilizada a matriz $\boldsymbol{u}_1\boldsymbol{q}_1'$. Em vez disso, substitui-se $\boldsymbol{u}_1$ pelo seu estimador $\widehat{\boldsymbol{u}}_1 = b_1\boldsymbol{t}_1$ para se obter a matriz de deflação $b_1\boldsymbol{t}_1\boldsymbol{q}_1'$.

Explicado de outra forma, dado que $\boldsymbol{Y} = \boldsymbol{UQ}' + \boldsymbol{F}$ e como $\boldsymbol{Y} = \boldsymbol{F}_0$, temos então que

$$\boldsymbol{F}_0 = \boldsymbol{u}_1\boldsymbol{q}_1' + \boldsymbol{F}.$$

Renomeando a primeira matriz de resíduos como $\boldsymbol{F}_1$ e substituindo $\boldsymbol{u}_1$ por seu estimador, então (Passo *2.2.4*):

$$\boldsymbol{F}_1 = \boldsymbol{F}_0 - b_1\boldsymbol{t}_1\boldsymbol{q}_1'.$$

3. Atualiza-se a matriz $\boldsymbol{B}$, computando o primeiro elemento da diagonal como b_1, e retorna-se ao passo 1 para repetir todo o procedimento com as matrizes deflacionadas $\boldsymbol{E}_1$ e $\boldsymbol{F}_1$. Assumindo-se que o posto de $\boldsymbol{X}$ é r, após a matriz dos preditores ter sido deflacionada na primeira iteração pela subtração de $\boldsymbol{t}_1\boldsymbol{p}_1'$, o seu posto passa a ser $(r-1)$. A cada iteração o posto é diminuído de 1, até que a matriz $\boldsymbol{X}$ depreciada se torne uma matriz nula. O mesmo não ocorre com $\boldsymbol{Y}$, pois como é utilizada uma estimativa para se aproximar $\boldsymbol{u}_1\boldsymbol{q}_1$ em sua deflação, o posto de $\boldsymbol{Y}$ não decresce de 1 a cada iteração.

   Portanto, ainda que r (o posto de $\boldsymbol{X}$) seja maior do que o posto de $\boldsymbol{Y}$, o algoritmo pode ser repetido até r iterações. O número máximo de componentes k a serem extraídas pelo algoritmo será então igual ao posto de $\boldsymbol{X}$, ou seja, $k \leq r$ e, ao final, tem-se que:

$$\hat{\boldsymbol{Y}} = \hat{b}_1\boldsymbol{T}_1 + \hat{b}_2\boldsymbol{T}_2 + \cdots + \hat{b}_k\boldsymbol{T}_k.$$

2.3.3 A predição de Y

O estimador pelo método OLS do parâmetro β do modelo de regressão linear múltipla $Y = X\beta$ é calculado pela seguinte operação matricial:

$$\widehat{\boldsymbol{B}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Isso proporciona que, naquele modelo, as variáveis dependentes sejam preditas por intermédio da seguinte operação:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{B}}_{OLS}.$$

Já pelo método PLS, como o modelo utiliza a matriz dos scores de $\boldsymbol{X}$ para predizer a variável dependente centrada Y, a equação acima então pode ser matricialmente descrita na forma:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y} = \boldsymbol{T}\widehat{\boldsymbol{B}}$$

$$\text{sendo } \widehat{\boldsymbol{B}} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}.$$

Por outro lado, por serem calculadas a partir de sucessivas matrizes deflacionadas, as colunas da matriz $\boldsymbol{W}$ não podem ser comparadas diretamente. Contudo, prova-se [23] que $\boldsymbol{W}$ está relacionada a uma certa matriz $\boldsymbol{R}$, tal que:

$$\boldsymbol{R} = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}.$$

Com essa transformação, os pesos passam a estar diretamente relacionados a $\boldsymbol{X}$ e, semelhantemente ao que é feito no Passo *2.1.3* do algoritmo NIPALS, a matriz dos *scores* $\boldsymbol{T}$ pode ser descrita por $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{R}$. Então, retornando-se a $\widehat{\boldsymbol{Y}}$ e fazendo a substituição, chega-se a

$$\widehat{\boldsymbol{Y}} = \boldsymbol{T}\widehat{\boldsymbol{B}} =$$

$$\boldsymbol{X}\boldsymbol{R}\widehat{\boldsymbol{B}}$$

e, consequentemente[4]:

$$\widehat{\boldsymbol{B}}_{PLS} = \boldsymbol{R}\widehat{\boldsymbol{B}} = \boldsymbol{R}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y} =$$

$$\boldsymbol{R}\boldsymbol{T}'\boldsymbol{Y} =$$

$$\boldsymbol{R}\boldsymbol{Q}'.$$

Portanto, segue-se que o modelo preditivo para as variáveis dependentes é determinado pelo método PLS por intermédio de

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}_{PLS}.$$

Observe-se por último que, como as matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ se referem a dados centrados, o vetor $\overline{\boldsymbol{y}}$ deve ser somado aos resultados obtidos pela relação acima para que se obtenha uma estimativa para os dados originais.

[4] Notar que $\boldsymbol{Q}' = \boldsymbol{T}'\boldsymbol{Y}$ (vide Passo *2.1.5* do algoritmo NIPALS PLS2).

2.4 Quantidade de componentes e qualidade da predição

O objetivo maior do método PLS é predizer o valor de variáveis dependentes face a novas amostras coletadas de variáveis explicativas da mesma população de interesse. A quantidade de variância explicada por cada T_j indica sua importância nessa predição [1]. Considerando que $\boldsymbol{X}$ tem posto r e que a quantidade de componentes do modelo será $k \leq r$, a variável latente deve ser extraída somente se ela for relevante para a melhoria da predição de $\boldsymbol{Y}$, sob pena de se incorrer em sobreajuste do modelo, ou seja, a retenção de componentes menores no modelo pode significar apenas a descrição de ruídos.

2.4.1 Validação cruzada e qualidade da predição

Para se estimar a quantidade de variáveis latentes que devem ser retidas no modelo quando o propósito é a predição, utiliza-se o procedimento de reamostragem conhecido como *validação cruzada*. Este procedimento consiste em se particionar o conjunto de dados, com n observações, em c subconjuntos mutuamente exclusivos, cada um com a observações. Caso n seja par, então a quantidade mínima de subconjuntos será $c = 2$, cada um com $a = n/2$ observações.

A quantidade máxima de subconjuntos será $c = n$ e o número de observações por subconjunto será de $a = 1$. Essa última configuração de validação cruzada é chamada de deixa-um-fora (*leave-one-out*), enquanto as demais são conhecidas na literatura como *k-fold*. Se n é ímpar, a única diferença é que um dos subconjuntos poderá ficar com uma observação a mais, dependendo da escolha de c.

Para um modelo em que serão extraídas k componentes PLS, a técnica de validação cruzada consiste em separar um desses subconjuntos como conjunto de teste e ajustar o modelo utilizando os $(c-1)$ conjuntos de treino restantes na estimação dos parâmetros. Em seguida, calcula-se as predições para o conjunto de teste $\hat{y}_i$ e a soma dos quadrados dos erros:

$$SQE = \sum_{i=1}^{a} (y_i - \hat{y}_i)^2.$$

Em seguida, é separado um outro subconjunto para servir como conjunto de teste, repetindo-se o procedimento até que cada um dos c subconjuntos tenha sido usado como conjunto de teste uma única vez. Depois de c iterações, calcula-se

$$PRESS_k = \sum_{j=1}^{c} SQE_j.$$

O *PRESS*, que é um acrônimo da expessão em inglês *predicted residual sum of squares*, é uma medida do poder de predição do modelo com k componentes. Quanto menor for o valor do PRESS, maior será a qualidade do ajuste. Quando adicionamos mais uma variável latente ao modelo e, com essas $(k+1)$ componentes, o ganho na capacidade de predição é negligenciável, então a quantidade ideal de fatores será igual a k. Na literatura, considera-se que a melhoria no poder de predição não será significativa se, com extração de mais uma componente, a seguinte razão se verificar [17]:

$$Q_{k+1} = \frac{PRESS_{k+1}}{PRESS_k} > 0{,}9.$$

Caso a qualidade de predição decresça com o aumento da quantidade de componentes, significa que o modelo está sobreajustado [1]. Um outro critério para avaliar a qualidade do modelo preditivo é utilizar a medida *RMSEP* (do inglês *Root Mean Squared Error of Prediction*), definida como:

$$RMSEP_k = \sqrt{PRESS_k/c}.$$

Por intermédio da representação gráfica da quantidade de componentes contra o RMSEP, é possível se utilizar o método do cotovelo [17] para se determinar quantas componentes devem ser retidas no modelo. Como exemplo do procedimento, considere-se o ficheiro *gasoline* [13] do pacote *pls* do R. Esse é um conjunto de dados que contém a octanagem e o espectro de absorção NIR de 60 amostras de gasolina, cujo gráfico é mostrado na Figura 2.2. Cada espectro NIR consiste em 401 medições de refletância difusa de 900 até 1700 nanômetros [4], ao qual será aplicado o modelo PLS1.

A estratégia a ser utilizada é o cálculo do PRESS e do RMSEP com a alteração do número de componentes retidas a cada iteração, ou seja, fazendo k variar até uma determinada quantidade de variáveis latentes previamente fixadas. Nesse caso, fez-se $k = 1, \ldots, 10$ e foi aplicado o algoritmo NIPALS PLS1 aos dados. Para cada k, as predições foram feitas utilizando-se a validação cruzada do tipo *deixa-um-fora* , em que $c = n = 60$ e $a = 1$.

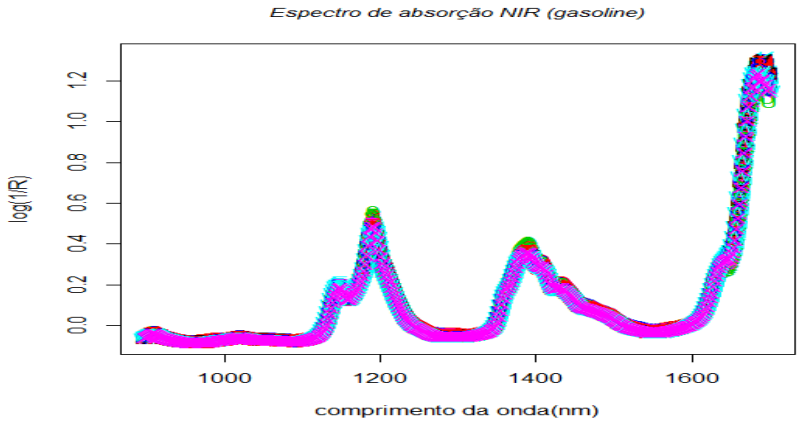

Figura 2.2: Espectro de absorção NIR (*Near Infrared Reflectance*) de 60 amostras de gasolina.

Pela Tabela 2.1, observa-se que a partir de $k = 3$ o benefício com o acréscimo de mais uma componente é pequeno. O que é confirmado visualmente pelo gráfico da Figura 2.3, onde é mostrado o número de componentes contra o PRESS e a formação do cotovelo quando utilizadas três variáveis latentes em sua construção. O mesmo ocorre na Figura 2.4, onde a mesma abordagem é utilizada, mas adotando-se o RMSEP como medida.

K	***PRESS***	***RMSEP***
1	105.842	1.328
2	8.724	0.381
3	3.991	0.258
4	3.489	0.241
5	3.489	0.241
6	3.159	0.229
7	2.881	0.219
8	3.118	0.228
9	3.519	0.242
10	3.574	0.244

Tabela 2.1: Determinação da quantidade ideal de componentes a serem retidas pelo NIPALS com base nos índices PRESS e RMSEP, que medem a falta de precisão do modelo na predição, sendo que quanto maior o índice menor é a sua acurácia.

Uma outra maneira de mostrar visualmente que $k = 3$ é a quantidade ideal de componentes PLS para os dados *gasoline* é através do gráfico representado na Figura 2.5, na qual os *valores observados* e os *valores preditos* são confrontados, para $k = 1, \ldots, 4$. No gráfico, o segmento em azul mostra a reta teórica quando se verifica $x = y$. É possível se observar que a qualidade da predição aumenta com o acréscimo de variáveis latentes até a terceira componente, mas da terceira para a quarta não há ganho perceptível.

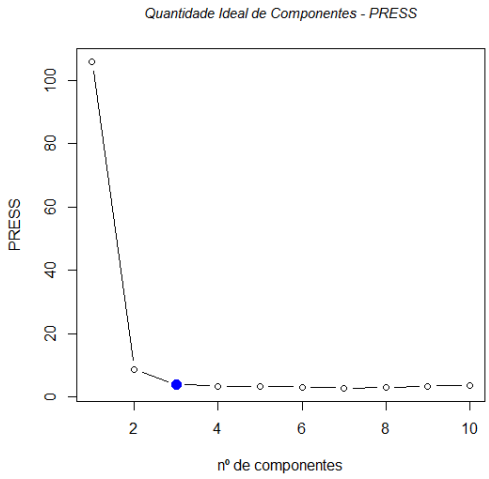

Figura 2.3: Representação gráfica da determinação da quantidade de variáveis latentes que devem ser utilizadas no modelo para evitar o sobreajuste. Verifica-se que a partir de $k = 3$ (em azul) a variação no PRESS é relativamente menor, formando o cotovelo. Isto sugere que a quantidade de componentes a serem retidas é igual a 3. (*Dados gasoline do pacote pls do R*)

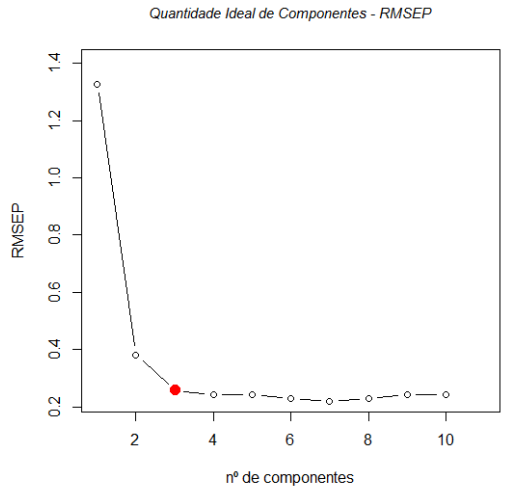

Figura 2.4: Determinação da quantidade de variáveis latentes utilizando o RMSE. (*Dados gasoline do pacote pls do R*)

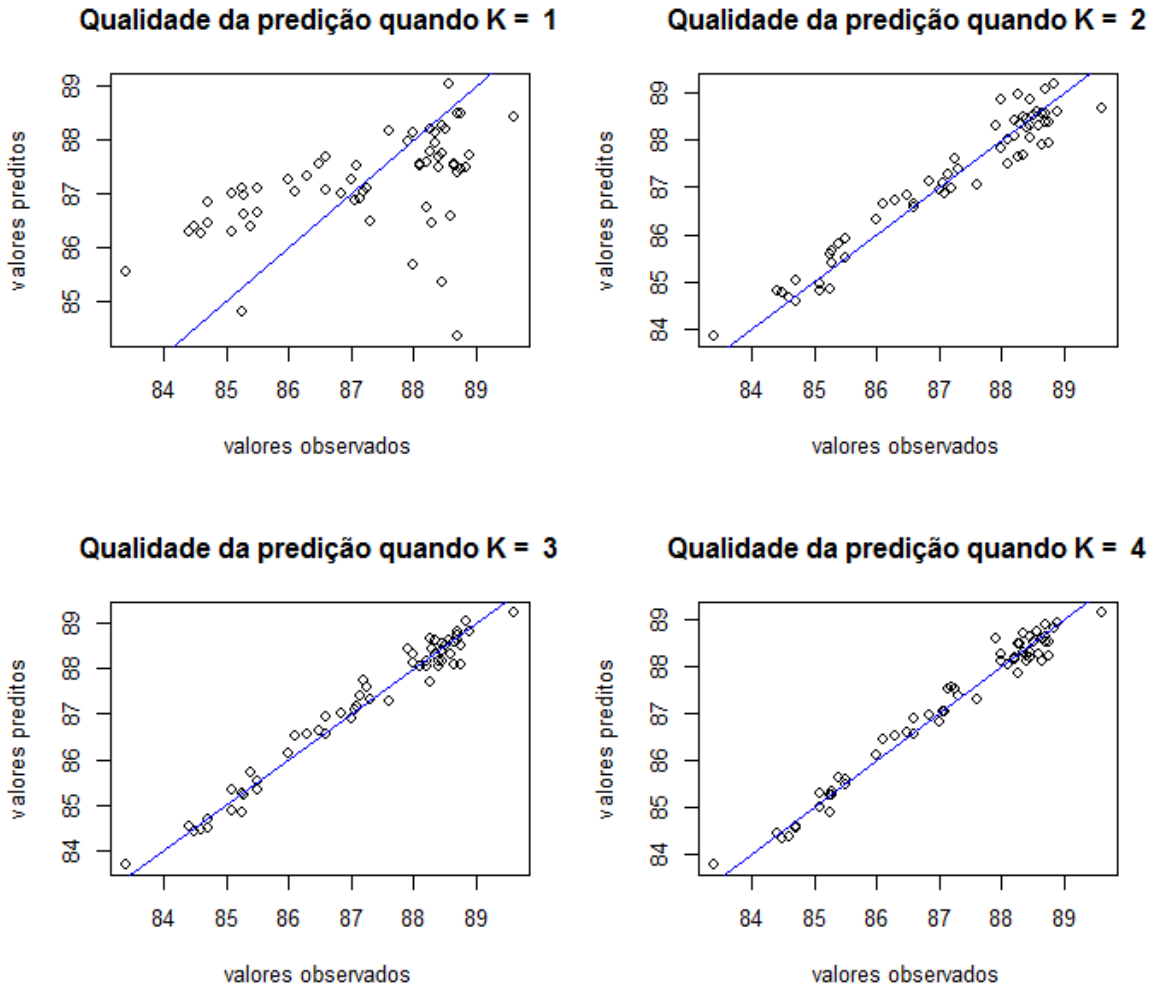

Figura 2.5: Validação cruzada do tipo deixa-um-fora, a predição feita quando são extraídas 3 componentes PLS pelo NIPALS se mostra a mais adequada, dado que não há melhora perceptível com o acréscimo de mais um fator, e sob pena de se incorrer em sobreajuste.

Capítulo 3

O *biplot* do PLS

3.1 Visualização de dados multivariados

O desenvolvimento de técnicas de exibição gráfica de dados com estrutura complexa é justificado pela habilidade possuída pelo cérebro humano de perceber, analisar e interpretar a informação visual [24]. A utilização de imagens como meio facilitador para a percepção dos padrões existentes entre as observações, assim como das relações entre variáveis acaba sendo uma vantagem adicional quando o investigador está diante de matrizes de dados com elevada dimensionalidade.

A utilização da visualização, como ferramenta exploratória de análise de resultados do PLS, permite a compreensão de conceitos importantes envolvidos no método, tais como *scores*, *loadings*, projeções das observações, direções das componentes e importância de cada variável latente na predição de $\boldsymbol{Y}$. O método de exibição gráfica que será tratado nesta dissertação é o *biplot* PLS, que é uma variante do *biplot* tradicional adaptada à metodologia PLS.

3.2 O *biplot* e o PLS

O termo *biplot* é devido a [8], que propôs a generalização dos diagramas de dispersão de dados bivariados para problemas multivariados, com a representação gráfica simultânea dos indivíduos (por pontos) e das variáveis (por vetores) de um conjunto de dados multivariados. A ideia é permitir a visualização gráfica de uma matriz de dados de forma resumida, mas de maneira clara e metódica o suficiente para expor a estrutura principal dessas informações, como por exemplo, padrões de correlações entre variáveis e similaridades entre observações [11].

A representação gráfica *biplot* é construída com base na decomposição de uma *matriz alvo* $\boldsymbol{D}$, de dimensão nxm, no produto de duas outras matrizes, de dimensões nxk e kxm, e que aqui serão chamadas de *matriz direita* e *matriz esquerda*, tal que:

$$\boldsymbol{D} = \boldsymbol{G}\boldsymbol{H}'.$$

A ideia básica por trás do *biplot* é que os elementos da matriz alvo $\boldsymbol{D}$ são iguais ao produtos escalares entre os correspondentes pares de vetores das linhas de $\boldsymbol{G}$ e $\boldsymbol{H}$. Pelas regras da multiplicação de matrizes, os elementos da linha i da matriz esquerda são multiplicados pelos correspondentes elementos da coluna j da matriz direita transposta. Em seguida, são adicionados para produzir o elemento a_{ij} da matriz alvo. Essa soma de produtos-cruzados é chamada de *produto escalar* e é a base da geometria do *biplot*. Tem-se então que, para quaisquer dois vetores $\mathbf{a}' = [a_1\ a_2 \dots a_m]$ e $\mathbf{b}' = [b_1\ b_2 \dots b_m]$, com m elementos cada, a definição algébrica do produto escalar entre $\mathbf{a}$ e $\mathbf{b}$ é

$$\mathbf{a}'\mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_mb_m.$$

Assim, considerando o ponto $\boldsymbol{g}_i$, a i-ésima linha de $\boldsymbol{G}$, que representa a i-ésima observação dos dados originais, e também o vetor $\boldsymbol{h}_j$, a j-ésima coluna de $\boldsymbol{H}'$, que representa a j-ésima variável, então a representação matricial da decomposição da matriz alvo $\boldsymbol{D}$ pode ser feita por intermédio dos produtos escalares que formam os seus elementos, tal que:

$$\boldsymbol{D} = \boldsymbol{G}\boldsymbol{H}' = \begin{pmatrix} g_1' \\ \vdots \\ g_n' \end{pmatrix} (h_1 \quad \dots \quad h_m) = \begin{pmatrix} g_1'h_1 & \cdots & g_1'h_m \\ \vdots & \ddots & \vdots \\ g_n'h_1 & \cdots & g_n'h_m \end{pmatrix}.$$

As matrizes $\boldsymbol{G}$ e $\boldsymbol{H}$ que surgem nesta decomposição formam dois conjuntos de pontos. Se estes pontos são bidimensionais (i.e., $k = 2$), então as linhas e colunas de $\boldsymbol{D}$ podem ser representadas por meio de um gráfico bidimensional, com os n pontos do primeiro conjunto ($\boldsymbol{G}$) representados por pontos e com os m pontos do segundo conjunto ($\boldsymbol{H}$) reproduzidos na forma de vetores conectados à origem. Se o ponto $\boldsymbol{g}_i$ for projetado sobre o eixo determinado por $\boldsymbol{h}_j$ e, em seguida, a norma dessa projeção for multiplicada pela pela norma de $\boldsymbol{h}_j$, o resultado será equivalente à definição geométrica[1] do produto escalar. Desta forma, é possível se representar o elemento d_{ij} da matriz alvo $\boldsymbol{D}$ tanto pela definição algébrica, quanto pela definição geométrica do produto escalar, tal que:

$$d_{ij} = \boldsymbol{g}_i'\,\boldsymbol{h}_j = \|\boldsymbol{g}_i\|\|\boldsymbol{h}_j\|cos\theta.$$

[1] O produto escalar entre x e y pode ser escrito como x.y = |x||y| cosθ, sendo θ o ângulo entre eles.

Na Figura 3.1, mostra-se a representação geométrica dos elementos envolvidos no cálculo do produto escalar, ou seja, os comprimentos dos vetores e a projeção escalar, que tem como resultado o valor de cada elemento d_{ij} da matriz alvo.

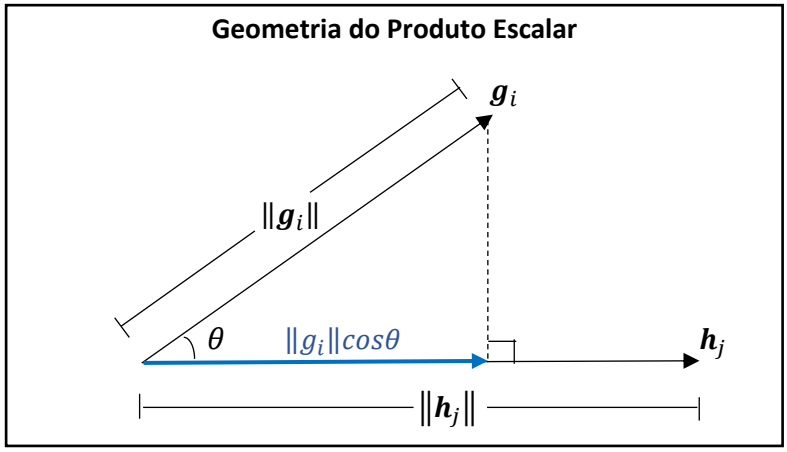

Figura 3.1: Representação geométrica dos elementos que originam os valores da matriz alvo, tal que $d_{ij} = \boldsymbol{g}_i' \boldsymbol{h}_j = \|\boldsymbol{g}_i\|\|\boldsymbol{h}_j\| cos\theta$ da matriz alvo.

Denomina-se *ponto biplot* cada conjunto de coordenadas formado por uma linha da matriz $\boldsymbol{G}$, e *vetor biplot* cada conjunto de coordenadas composto por uma linha de $\boldsymbol{H}$. A quantidade de coordenadas desses pontos ou vetores, traduzida pelo número de colunas de $\boldsymbol{G}$ e $\boldsymbol{H}$, é determinada pelo posto da matriz alvo. Portanto, para que se possa representar uma matriz de dados com elevada dimensionalidade em gráficos *biplot* bidimensionais ou tridimensionais, torna-se necessário a aplicação de algum método que, ao mesmo tempo, decomponha e encontre uma aproximação da matriz alvo em um espaço mais reduzido.

Diante de uma matriz de dados com muitas variáveis e posto r, o que se busca é aproximar essa matriz por uma outra com posto s, tal que $s < r$, mas que se assemelhe à primeira tão proximamente quanto for possível. No PLS, o conjunto de dados diz respeito a duas matrizes centradas: $\boldsymbol{X}_{nxm}$ e $\boldsymbol{Y}_{nxp}$, que representam, respectivamente, as variáveis explicativas e as variáveis respostas. Portanto, considerando que a matriz alvo seja formada por esses dois blocos, $\boldsymbol{X}$ e $\boldsymbol{Y}$, então $\boldsymbol{D}$ terá dimensão n x $(m + p)$, assim representada:

$$\boldsymbol{D} = [\boldsymbol{X} \quad \boldsymbol{Y}].$$

No PLS, sabe-se que é possível se conseguir uma aproximação $\widetilde{\boldsymbol{X}}$ para $\boldsymbol{X}$ por meio do produto de sua matriz de *scores* pela sua matriz de *loadings*. De maneira similar ao que é feito no Passo 2.2 do NIPALS PLS2 (regressão de $\boldsymbol{X}$ sobre $\boldsymbol{T}$), a aproximação é obtida por:

$$\widetilde{\boldsymbol{X}} = \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{X} = \boldsymbol{T}\boldsymbol{T}'\boldsymbol{X} \cong \boldsymbol{T}\boldsymbol{P}'$$

$$\text{onde } \boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}, \text{a matriz identidade.}$$

Analogamente, é possível a obtenção de uma matriz de aproximação $\widetilde{\boldsymbol{Y}}$ para $\boldsymbol{Y}$ utilizando-se $\boldsymbol{T}$ em vez de $\boldsymbol{U}$ [17]. Para tal, faz-se a regressão de $\boldsymbol{Y}$ sobre $\boldsymbol{T}$ e, por intermédio da relação estabelecida no Passo *2.1.5* do NIPALS PLS2, chega-se a:

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{T}'\boldsymbol{Y} \cong \boldsymbol{T}\boldsymbol{Q}'.$$

A matriz $\boldsymbol{Q}$ é formada pelos vetores de pesos de $\boldsymbol{Y}$, conforme mostrado na Seção 2.3. Portanto, a matriz de aproximação para $\boldsymbol{D}$ é obtida de forma que:

$$\widetilde{\boldsymbol{D}} = \left[\, \widetilde{\boldsymbol{X}} \quad \widetilde{\boldsymbol{Y}} \,\right] = [\; \boldsymbol{T}\boldsymbol{P}' \quad \boldsymbol{T}\boldsymbol{Q}'] = \boldsymbol{T}[\, \boldsymbol{P} \quad \boldsymbol{Q} \,]'.$$

Extraindo-se apenas duas variáveis latentes, a matriz $\boldsymbol{T}$ dos scores terá dimensão $n\text{x}2$ e a matriz formada pelos blocos $\boldsymbol{P}$ e $\boldsymbol{Q}$ terá dimensão $2\text{x}(m+p)$. As linhas de $\boldsymbol{T}$ são definidas então como os pontos *biplot* e representam as observações da amostra, enquanto que as colunas da matriz $[\, \boldsymbol{P} \quad \boldsymbol{Q} \,]'$ determinam os vetores *biplot* e retratam as variáveis, sendo que as colunas de 1 até m se referem às variáveis explicativas e as colunas de $(m+1)$ até $(m+p)$ dizem respeito às variáveis dependentes. Por fim, as extensões dos vetores *biplot* formam os eixos *biplot*, sobre os quais os pontos *biplot* podem ser projetados para se estimar os elementos da matriz de aproximação.

A título de exemplificação de construção de um *biplot* PLS, considere-se o ficheiro *oliveoil* do pacote *pls* do R [4]. Os dados contidos se referem a dezesseis amostras de azeite, sendo que as cinco primeiras observações são de produtos gregos, as cinco seguintes são óleos de oliva italianos e as seis últimas são de azeites espanhóis. As variáveis explicativas representam cinco parâmetros de qualidade de natureza físico-química e aparecem na Tabela 3.1, formando a matriz $\boldsymbol{X}$ do PLS.

	Acidity	***Peroxide***	***K232***	***K270***	***DK***
G1	0.73	12.7	1.9	0.139	0.003
G2	0.19	12.3	1.678	0.116	-0.004
G3	0.26	10.3	1.629	0.116	-0.005
G4	0.67	13.7	1.701	0.168	-0.002
G5	0.52	11.2	1.539	0.119	-0.001
I1	0.26	18.7	2.117	0.142	0.001
I2	0.24	15.3	1.891	0.116	0
I3	0.3	18.5	1.908	0.125	0.001
I4	0.35	15.6	1.824	0.104	0
I5	0.19	19.4	2.222	0.158	-0.003
S1	0.15	10.5	1.522	0.116	-0.004
S2	0.16	8.14	1.527	0.1063	-0.002
S3	0.27	12.5	1.555	0.093	-0.002
S4	0.16	11	1.573	0.094	-0.003
S5	0.24	10.8	1.331	0.085	-0.003
S6	0.3	11.4	1.415	0.093	-0.004

Tabela 3.1: Matriz X - Variáveis explicativas do ficheiro *oliveoil*, pacote *pls* do R.

As variáveis dependentes estão associadas a seis atributos verificados em análise sensorial e formam a matriz $\boldsymbol{Y}$**,** cuja composição é mostrada na Tabela 3.2:

	Yellow	***Green***	***Brown***	***Glossy***	***Transp***	***Syrup***
G1	21.4	73.4	10.1	79.7	75.2	50.3
G2	23.4	66.3	9.8	77.8	68.7	51.7
G3	32.7	53.5	8.7	82.3	83.2	45.4
G4	30.2	58.3	12.2	81.1	77.1	47.8
G5	51.8	32.5	8	72.4	65.3	46.5
I1	40.7	42.9	20.1	67.7	63.5	52.2
I2	53.8	30.4	11.5	77.8	77.3	45.2
I3	26.4	66.5	14.2	78.7	74.6	51.8
I4	65.7	12.1	10.3	81.6	79.6	48.3
I5	45	31.9	28.4	75.7	72.9	52.8
S1	70.9	12.2	10.8	87.7	88.1	44.5
S2	73.5	9.7	8.3	89.9	89.7	42.3
S3	68.1	12	10.8	78.4	75.1	46.4
S4	67.6	13.9	11.9	84.6	83.8	48.5
S5	71.4	10.6	10.8	88.1	88.5	46.7
S6	71.4	10	11.4	89.5	88.5	47.2

Tabela 3.2: Matriz Y - Variáveis dependentes do ficheiro *oliveoil*, pacote *pls* do R.

Após a normalização da base de dados *oliveoil*, foi aplicado o método PLS com a utilização do algoritmo NIPALS para a extração de duas componentes, tal que $\widetilde{\boldsymbol{X}}_{16\text{x}5} = \boldsymbol{T}_{16\text{x}2}\boldsymbol{P}'_{2\text{x}5}$ e $\widetilde{\boldsymbol{Y}}_{16\text{x}6} = \boldsymbol{T}_{16\text{x}2}\boldsymbol{Q}'_{2\text{x}6}$. Os resultados são mostrados na Tabela 3.3, com as matrizes de *loadings* de $\boldsymbol{P}$ e $\boldsymbol{Q}$, e na Tabela 3.4, com os dados relativos aos *scores* que compõem a matriz $\boldsymbol{T}$.

	P1	***P2***
Acidity	-0.24	-0.81
Peroxide	-0.50	0.36
K232	-0.54	0.27
K270	-0.48	-0.13
DK	-0.39	-0.34
	Q1	***Q2***
yellow	0.37	0.20
green	-0.34	-0.26
brown	-0.38	0.52
glossy	0.42	-0.07
transp	0.39	0.01
syrup	-0.40	0.21

Tabela 3.3: Matrizes $\boldsymbol{P}$ e $\boldsymbol{Q}$ - *loadings* com extração de 2 componentes (*Oliveoil*).

	T1	*T2*
G1	-1.95	-2.50
G2	0.72	0.70
G3	1.20	0.21
G4	-1.51	-1.73
G5	0.33	-1.48
I1	-2.62	0.87
I2	-0.84	0.54
I3	-1.80	0.55
I4	-0.62	0.11
I5	-2.67	1.89
S1	1.41	0.45
S2	1.70	-0.16
S3	1.08	0.09
S4	1.53	0.51
S5	2.21	-0.05
S6	1.81	-0.03

Tabela 3.4: Duas componentes PLS extraídas de X da base de dados *oliveoil*.

Essas são as matrizes que permitem não somente calcular a matriz de aproximação $\widetilde{\boldsymbol{D}}$, mas também construir o *biplot* PLS associado, mostrado na Figura 3.2. Os pontos em castanho se referem às observações e refletem as linhas da matriz $\boldsymbol{T}$. Em vermelho, as linhas de $\boldsymbol{P}$ estão representadas por vetores e representam os eixos *biplot* relacionados com os preditores e, em azul, as linhas de $\boldsymbol{Q}$ são retratadas por vetores e representam os eixos *biplot* associados às variáveis dependentes.

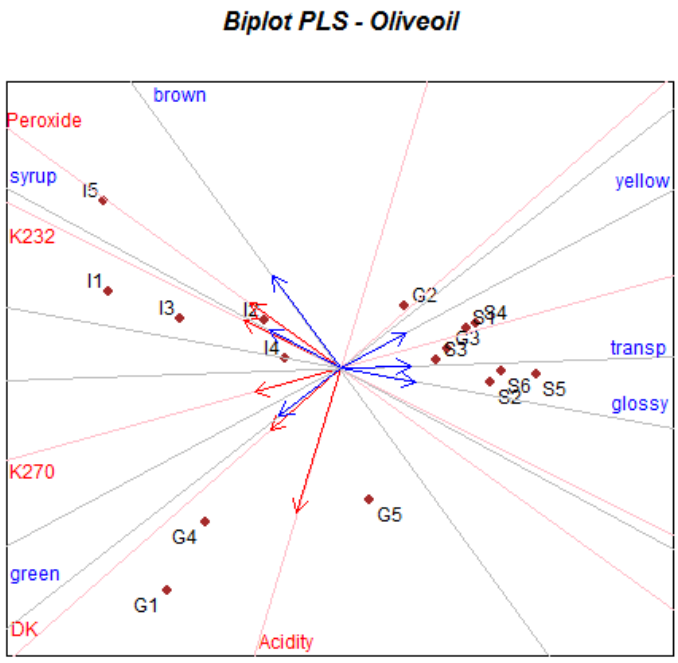

Figura 3.2: *Biplot* PLS aplicado - dados *Oliveoil* do pacote *pls* do R

3.3 Calibragem dos eixos *biplot*

O *biplot* do PLS pode ser aprimorado para que seja possível predizer os valores dos elementos das matrizes de aproximação $\widetilde{\boldsymbol{X}}$ e $\widetilde{\boldsymbol{Y}}$ diretamente do gráfico. Isso é feito por meio da calibragem e consiste em se traçar marcas ao longo dos eixos *biplot*, atribuindo-se valores de tal forma que a projeção das observações sobre a direção dos vetores *biplot* forneçam um valor aproximado de $\widetilde{\boldsymbol{x}}$ ou $\widetilde{\boldsymbol{y}}$.

Considerando que o objetivo seja calibrar o eixo *biplot* associado à j-ésima coluna da matriz centrada $\boldsymbol{X}$, o que se pretende é obter uma estimativa para $\widetilde{\boldsymbol{x}} = (\tilde{x}_{1j},\ \tilde{x}_{2j}, \dots ,\ \tilde{x}_{nj})$, ou seja, a j-ésima coluna da matriz de aproximação $\widetilde{\boldsymbol{X}}$, diretamente da projeção de cada observação $\boldsymbol{t}_i'$, $i = 1, \dots, n$, sobre a direção de $\boldsymbol{p}_j$, a j-ésima coluna de $\boldsymbol{P}'$. Assim, para um j fixo:

$$\widetilde{\boldsymbol{x}}_j = \begin{pmatrix} \boldsymbol{t}_1' \\ \vdots \\ \boldsymbol{t}_n' \end{pmatrix} \boldsymbol{p}_j.$$

Posto que $\boldsymbol{t}'\boldsymbol{p} = \|\boldsymbol{t}\|\|\boldsymbol{p}\| cos\theta_{\boldsymbol{t},\boldsymbol{p}}$, então a projeção de cada observação (linhas de $\boldsymbol{T} = \boldsymbol{t}_i'$) sobre o vetor *biplot* $\boldsymbol{p}_j$ será dado por

$$\|\boldsymbol{t}_i\| cos\theta_{\boldsymbol{t},\boldsymbol{p}} = \frac{\boldsymbol{t}_i'\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}.$$

Como os vetores $\boldsymbol{t}_i'$ e $\boldsymbol{p}_j$ são dados pelo NIPALS, basta calcular $\frac{\boldsymbol{t}_i'\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|}$. Esses são os pontos, na escala do gráfico, onde serão indicados os marcadores nos eixos *biplot* para cada uma das projeções das observações. A cada um desses marcadores de escala, assinala-se o correspondente valor[2] predito acrescido da média da coluna j, ou seja, $\widetilde{\boldsymbol{x}}_j + \bar{x}_j$ (pois os dados foram centrados). Subtituindo-se $\boldsymbol{p}_j$ por $\boldsymbol{q}_j$, chega-se aos marcadores de escala dos eixos *biplot* das variáveis dependentes, bastando associá-los ao respectivo $\widetilde{\boldsymbol{y}}_j + \bar{y}_j$.

Utilizando novamente a base dados *Oliveoil*, será determinado o primeiro marcador de escala na calibragem do eixo *biplot* referente à variável explicativa *Acidity*. Primeiramente, verifica-se na Tabela 3.4 as coordenadas da primeira observação *biplot*, que é, neste caso, o vetor $\boldsymbol{t}_1' = (-1.95,\ -2.50)$. Em seguida, recorre-se à Tabela 3.3 para se obter o vetor dos *loadings* relativo à variável dependente *Acidity* (primeira linha da matriz $\boldsymbol{P}$), que é $\boldsymbol{p}_1 = (-0.24,\ -0.81)'$ e, então, calcula-se $\boldsymbol{t}_1'\boldsymbol{p}_1 = \tilde{x}_{11} \cong 2.51$ e $\|\boldsymbol{p}_1\| \cong 0.84$.

[2] Ou $\widetilde{\boldsymbol{x}}_j dp_j + \bar{x}_j$ no caso de dados normalizados, onde $\bar{x}_j$ é a média e dp_j é o desvio padrão da coluna *j*.

O primeiro marcador de escala para a variável *Acidity* será então calculado como:

$$\frac{t_1' p_1}{\|p_1\|} \cong 2.98.$$

Este é o tamanho da projeção de $\boldsymbol{t}_1'$ sobre $\boldsymbol{p}_1$ na escala do gráfico e que deve corresponder ao valor da aproximação $\tilde{x}_{11}$. Contudo, antes de fazer a atribuição de $\tilde{x}_{11}$ ao marcador de escala, é preciso ajustá-lo pelo desvio padrão e pela média da variável[3], fazendo $\tilde{x}_{11}\ dp_1 + \bar{x}_1$ (onde dp_1 é o desvio padrão e $\bar{x}_1$ é a média da variável *Acidity*), o que em termos concretos equivale a $(2.51)(0.18) + 0.31 \cong 0.76$. Na Figura 3.3, por meio do pacote *calibrate* do R, é mostrado o eixo Acidity já calibrado e o quão próximo uma estimativa feita diretamente no eixo *biplot* pode estar do valor calculado, uma vez que a leitura da projeção da observação G1 sobre o eixo *biplot* Acidity é visualmente próximo a 0.76.

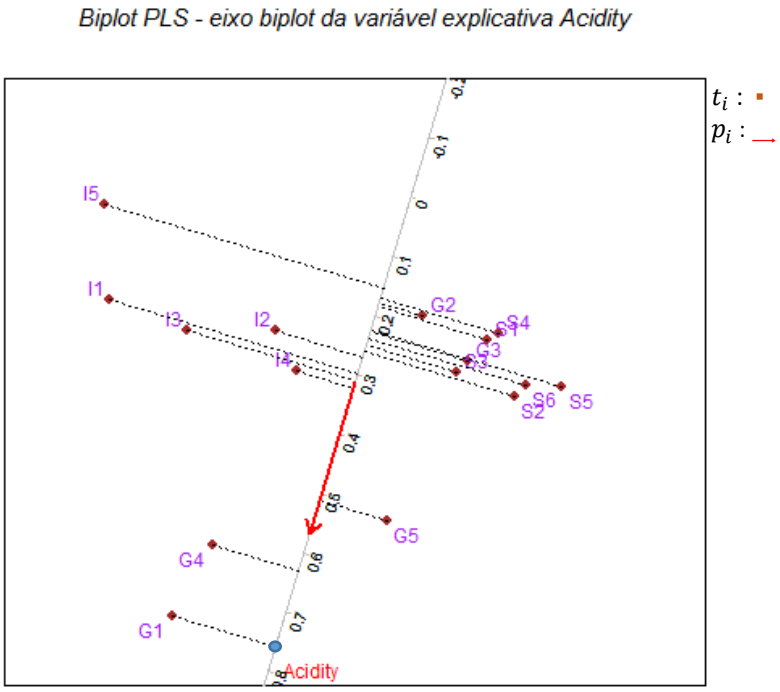

Figura 3.3: Eixo *biplot* da variável *Acidity* da base de dados Oliveoil calibrado. Em azul, a estimativa $\tilde{x}_{11}$obtida diretamente no gráfico é muito próxima ao valor calculado 0.76. Em vermelho, o vetor *biplot* da variável *Acidity*.

Em seguida, para fins de comparação, é mostrado na Tabela 3.5 e Tabela 3.6 as matrizes de aproximação $\widetilde{\boldsymbol{X}}$ e $\widetilde{\boldsymbol{Y}}$, já devidamente corrigidas pelo desvio padrão e média das colunas, e os respectivos gráficos *biplot* PLS calibrados na Figura 3.4 e Figura 3.5.

[3] Pois os dados foram normalizados. Além disso, em relação à matriz X (Tabela 3.1), a média e o desvio padrão de sua primeira coluna são $\bar{x}_1 \cong 0.31$ e $dp_1 \cong 0.18$.

	Acidity	*Peroxide*	*K232*	*K270*	*DK*
G1	0.76	13.49	1.8	0.149	0.002
G2	0.18	12.89	1.66	0.108	-0.003
G3	0.23	11.47	1.56	0.104	-0.003
G4	0.63	13.69	1.8	0.141	0.001
G5	0.51	10.86	1.56	0.119	-0.001
I1	0.3	18.79	2.12	0.146	0
I2	0.27	15.37	1.86	0.126	-0.001
I3	0.31	17.01	1.99	0.137	-0.001
I4	0.32	14.46	1.8	0.125	-0.001
I5	0.15	20.13	2.2	0.143	-0.001
S1	0.19	11.4	1.55	0.1	-0.003
S2	0.26	10.16	1.47	0.099	-0.003
S3	0.25	11.52	1.57	0.105	-0.003
S4	0.17	11.27	1.53	0.099	-0.004
S5	0.22	9.42	1.4	0.093	-0.004
S6	0.24	10.11	1.46	0.097	-0.003

Tabela 3.5: Matriz de aproximação $\tilde{\boldsymbol{X}} = \boldsymbol{TP}'$ já ajustada pela média e desvio padrão das variáveis.

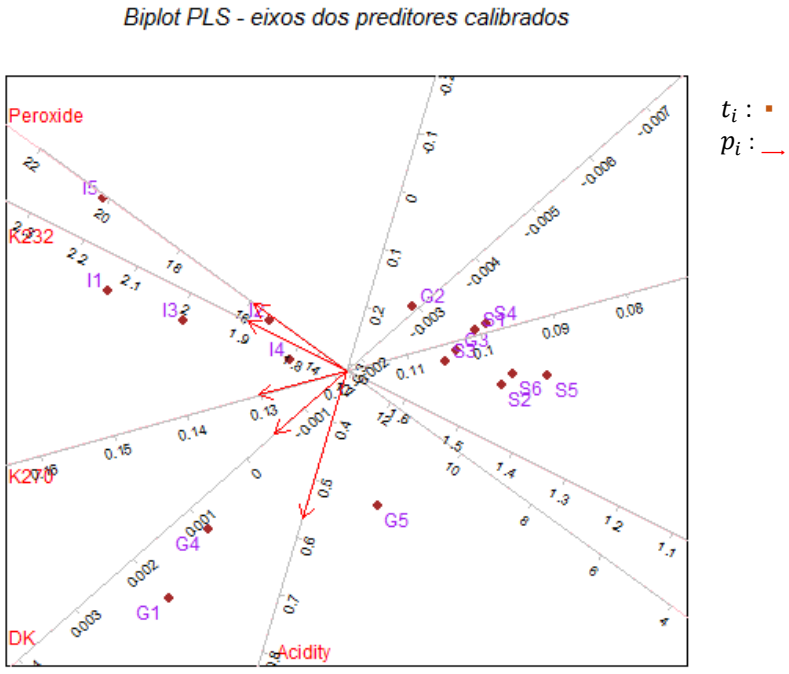

Figura 3.4: Eixos das variáveis preditoras do *biplot* PLS da base dados *Oliveoil* calibrados.

	yellow	*green*	*brown*	*glossy*	*transp*	*syrup*
G1	26.8	65.1	9.4	76.9	71.5	48.7
G2	58.9	23.2	12.8	82.4	80.7	47.5
G3	60.6	22.4	10.6	83.9	82.2	46.6
G4	33	56.7	10.6	77.7	73	48.7
G5	47.5	40.1	7.7	82.4	79.1	46.6
I1	35.1	49.2	19.8	73.6	69.7	51.8
I2	46.8	37	15.4	78.3	75.5	49.4
I3	39.9	44.6	17.3	75.8	72.3	50.6
I4	46.8	37.8	13.9	79.1	76.2	48.8
I5	38.7	43.2	22.6	72.9	69.6	52.6
S1	63	19.2	10.8	84.3	82.9	46.5
S2	62.7	20.8	8.6	85.3	83.8	45.8
S3	59.2	24.1	10.5	83.6	81.8	46.7
S4	64.1	17.8	10.7	84.6	83.3	46.4
S5	66.8	16	7.9	86.6	85.4	45.2
S6	64	19	8.7	85.6	84.2	45.7

Tabela 3.6: Matriz de aproximação $\widetilde{\boldsymbol{Y}} = \boldsymbol{TQ}'$ já ajustada pela média e desvio padrão das variáveis.

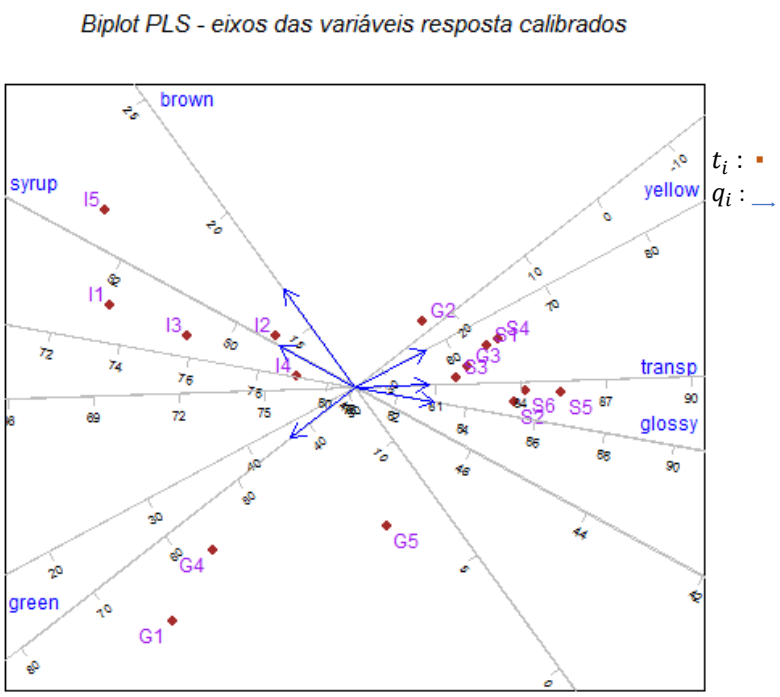

Figura 3.5: Eixos das variáveis resposta do *biplot* PLS da base dados *Oliveoil* calibrados.

Verifica-se, portanto, que a aproximação realizada por intermédio das projeções dos pontos sobre os eixos *biplot* calibrados fornece boas estimativas para os valores calculados de $\widetilde{\boldsymbol{X}}$ e $\widetilde{\boldsymbol{Y}}$.

3.4 Estimação dos coeficientes no *biplot*

Na Seção 2.3.2. foi visto que a matriz dos estimadores dos parâmetros PLS é definida como:

$$\widehat{\boldsymbol{B}}_{PLS} = \boldsymbol{R}\boldsymbol{Q}'.$$

Em que $\boldsymbol{R}$ é a matriz de pesos ajustados ao espaço da matriz $\boldsymbol{X}$ e determinada por $\boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}$, enquanto $\boldsymbol{Q}$ é a matriz dos *loadings* associada às variáveis dependentes. Atribuindo às linhas de $\boldsymbol{R}$ a condição de pontos *biplot* e estabelecendo-se as colunas de $\boldsymbol{Q}'$ como vetores *biplot*, constrói-se o *biplot* PLS da matriz dos coeficientes de regressão.

Após a construção do *biplot* PLS, considerando especificamente o eixo *biplot* de uma determinada variável dependente Y_j, a sua calibração permite que os coeficientes de regressão PLS associados sejam estimados diretamente no gráfico. Assim, para um *j* fixo, o vetor dos parâmetros referentes à Y_j é dado por:

$$\widehat{\boldsymbol{b}}_{PLS_j} = \begin{pmatrix} \boldsymbol{r}_1' \\ \vdots \\ \boldsymbol{r}_m' \end{pmatrix} \boldsymbol{q_j}.$$

Por conseguinte, os elementos do vetor $\widehat{\boldsymbol{b}}_{PLS_j}$ são obtidos por meio de produtos escalares, tal que:

$$\hat{b}_{PLS_{ij}} = \boldsymbol{r}_i' \, \boldsymbol{q}_j = \|\boldsymbol{r_i}\| \|\boldsymbol{q}_j\| cos\theta_{r_i,q_j} \, , i = 1, \dots, m.$$

Cada marcador de escala da projeção de um $\boldsymbol{r}_i$ sobre a direção de um vetor $\boldsymbol{q}_j$ é definido por $\boldsymbol{r}_i' \, \boldsymbol{q}_j \, / \, \|\boldsymbol{q}_j\| \, , \; i = 1, \dots, m$. A calibragem do eixo *biplot* de Y_j é finalizada com o cálculo de $\boldsymbol{r}_i' \, \boldsymbol{q}_j$, que é então atribuído ao marcador respectivo.

Para exemplificar a estimação dos coeficientes de regressão PLS por meio do *biplot* com eixos calibrados, retoma-se o resultado do PLS dos dados *Oliveoil*. As matrizes de interesse para o *biplot* PLS são mostradas na Tabela 3.7 , que traz os parâmetros da regressão PLS, na Tabela 3.8, que contém a matriz $\boldsymbol{R}$ dos pesos ajustados e a já referida Tabela 3.3, com os *loadings*.

	yellow	***green***	***brown***	***glossy***	***transp***	***Syrup***
Acidity	-0.23	0.28	-0.31	-0.03	-0.09	-0.08
Peroxide	-0.11	0.06	0.45	-0.26	-0.20	0.32
K232	-0.16	0.12	0.35	-0.26	-0.22	0.29
K270	-0.22	0.21	0.12	-0.20	-0.20	0.17
DK	-0.18	0.20	-0.06	-0.10	-0.13	0.05

Tabela 3.7: Matriz dos coeficientes de regressão PLS: $\boldsymbol{B}_{PLS}$ do *Oliveoil.*

	R1	***R2***
Acidity	-0.22	-0.76
Peroxide	-0.55	0.48
K232	-0.56	0.26
K270	-0.50	-0.14
DK	-0.31	-0.34

Tabela 3.8: Matriz $\boldsymbol{R}$, com os pesos ajustados.

Cada linha $\boldsymbol{r}_i'$ da matriz $\boldsymbol{R}$ (Tabela 3.8) é assinalada como um ponto no gráfico *biplot* (Figura 3.6), enquanto as linhas $\boldsymbol{q}_i$ da matriz $\boldsymbol{Q}$ (Tabela 3.3) são demarcadas como os vetores *biplot*. Por seu turno, esses últimos definem os eixos *biplot*, que são calibrados conforme mostrado anteriormente. Leituras de estimativas dos coeficientes de regressão PLS podem ser obtidas diretamente desse *biplot* ao se projetar ortogonalmente os pontos sobre os eixos *biplot*.

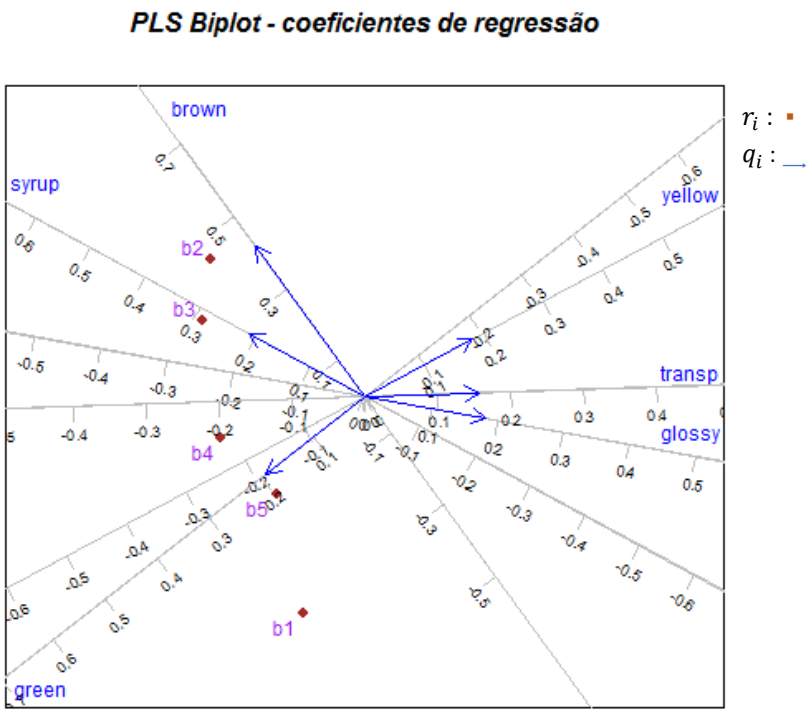

Figura 3.6: *Biplot* PLS dos Coeficientes de Regressão - *Oliveoil*.

Considerando a variável dependente *Yellow*, a Figura 3.7 mostra as projeções dos pesos r_i sobre o eixo *biplot* calibrado relativo àquela variável. Os pontos *biplot* receberam as etiquetas $b_1, \ldots, b_5$ e, neste caso específico, fazem referência à primeira coluna da Tabela 3.7, sendo b_1 o parâmetro associado ao preditor *Acidity*, b_2 ao *Peroxide* e assim por diante. Verifica-se que os valores obtidos com as projeções sobre o eixo calibrado são compatíveis com aqueles estimados e tabelados.

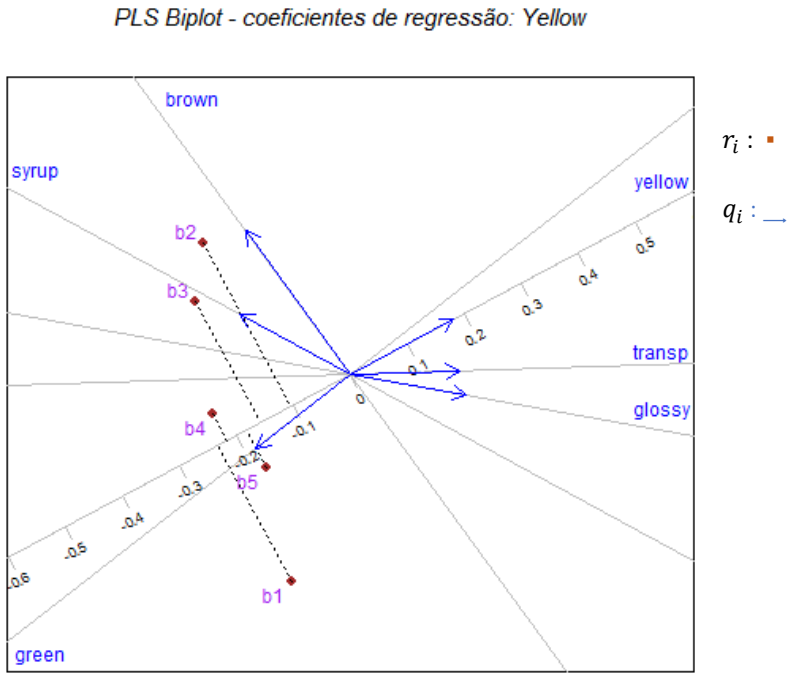

Figura 3.7: Leitura de estimativas dos coeficientes de regressão relativos à variável dependente Yellow diretamente no *Biplot* PLS – *Oliveoil*, obtidas com a projeção ortogonal de cada ponto *biplot* sobre o eixo *biplot* respectivo.

O procedimento é repetido para a variável dependente Brown e o resultado é mostrado na Figura 3.8, sendo que as leituras obtidas devem ser comparadas com a terceira coluna da Tabela 3.7:

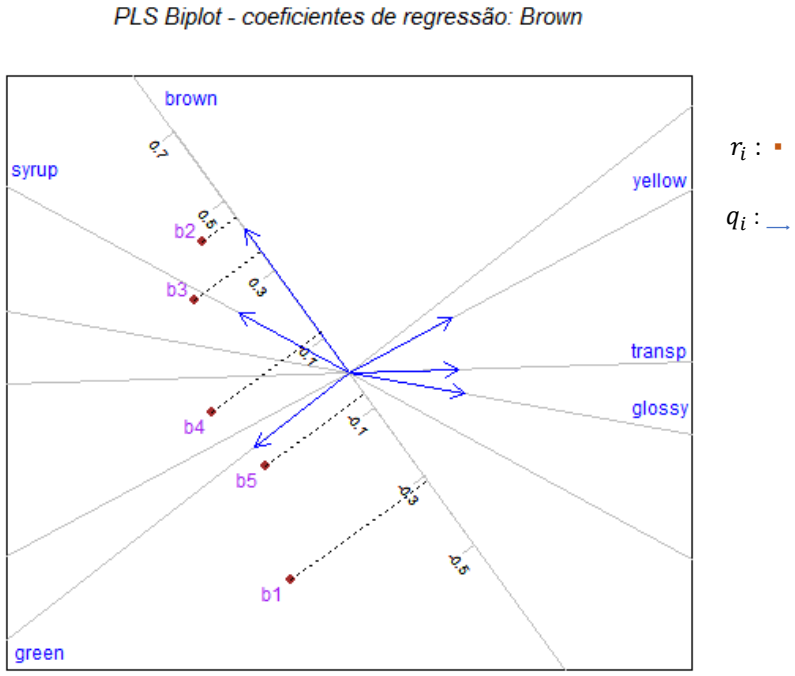

Figura 3.8: Estimativas dos coeficientes de regressão da variável dependente Brown.

3.5 Área *biplot*

No que se refere à obtenção e interpretação dos coeficientes de regressão $\widehat{\boldsymbol{b}}_{PLS}$ diretamente no gráfico *biplot*, [17] desenvolveu um método aplicável ao PLS. Nomeado como Área *Biplot*, esse método consiste em se obter estimativas para os coeficientes de regressão de forma visual e sem a necessidade de se proceder à calibragem dos eixos *biplot*.

Conforme se verifica no exemplo dos dados *Oliveoil*, Tabela 3.7, a matriz dos coeficientes de regressão $\boldsymbol{B}_{PLS}$ possui dimensão mxp, sendo seus elementos b_{ij}, $i = 1, \dots, m$ e $j = 1, \dots, p$, iguais a $\boldsymbol{r}'\boldsymbol{q}$. A ideia é rotacionar em 90° os pontos *biplot*, multiplicando o vetor $\boldsymbol{r}_i$ pela matriz de rotação

$$\Phi_{90^\circ} = \begin{pmatrix} cos\,90^\circ & sen\,90^\circ \\ -sen\,90^\circ & cos\,90^\circ \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Então, estimam-se as áreas dos triângulos formados pelos pontos rotacionados $\boldsymbol{r}_i^{\Phi} = \boldsymbol{r}_i\Phi_{90^\circ}$, a origem e os q_j. A importância dos preditores no modelo pode então ser avaliada, de forma visual, a partir da comparação das áreas de triângulos, bem como de suas posições em relação ao eixo *biplot*.

Considerando a i-ésma linha da matriz $\boldsymbol{R}$ rotacionada em 90° e a j-ésima coluna de $\boldsymbol{Q}'$, e assumindo que o ângulo entre r_i e q_j é θ_{ij}, então o ângulo formado entre r_i^{Φ} e q_j será $90^\circ + \theta_{ij}$ ou $90^\circ - \theta_{ij}$, dependendo se, em relação a q_j, o vetor r_i está ou não mais próximo de 0°. Em qualquer caso,

$$cos\,\theta_{ij} = sen\big(90^\circ + \theta_{ij}\big) = sen\big(90^\circ - \theta_{ij}\big).$$

Além disso, a matriz de rotação Φ_{90° preserva a norma de r_i, tal que:

$$\left\| r_i \right\| = \left\| r_i^{\Phi} \right\|.$$

Consequentemente, as seguintes igualdades são verificadas e, conforme pode ser observado na Figura 3.9, o segmento $\|r_i^{\Phi}\|sen\big(90^\circ - \theta_{ij}\big)$ é a altura do triângulo $A\hat{B}C$, cujos vértices são a origem, $\boldsymbol{r}_i^{\Phi}$ e $\boldsymbol{q}_j$:

$$\boldsymbol{r}_i'\boldsymbol{q}_j = \left\| \boldsymbol{r}_i \right\| \left\| \boldsymbol{q}_j \right\| cos\theta_{ij} = \left\| \boldsymbol{r}_i^{\Phi} \right\| \left\| \boldsymbol{q}_j \right\| sen\big(90^\circ - \theta_{ij}\big).$$

Definindo S_{ij} como a área do triângulo $A\hat{B}C$, então esta será calculada por:

$$S_{ij} = \frac{\left\| \boldsymbol{r}_i^{\Phi} \right\| \left\| \boldsymbol{q}_j \right\| sen(90° - \theta_{ij})}{2},$$

fazendo com que se estabeleça a seguinte relação entre o coeficiente de regressão b_{ij} e a área do do triângulo $A\hat{B}C$:

$$b_{ij} = 2\, x\, S_{ij}.$$

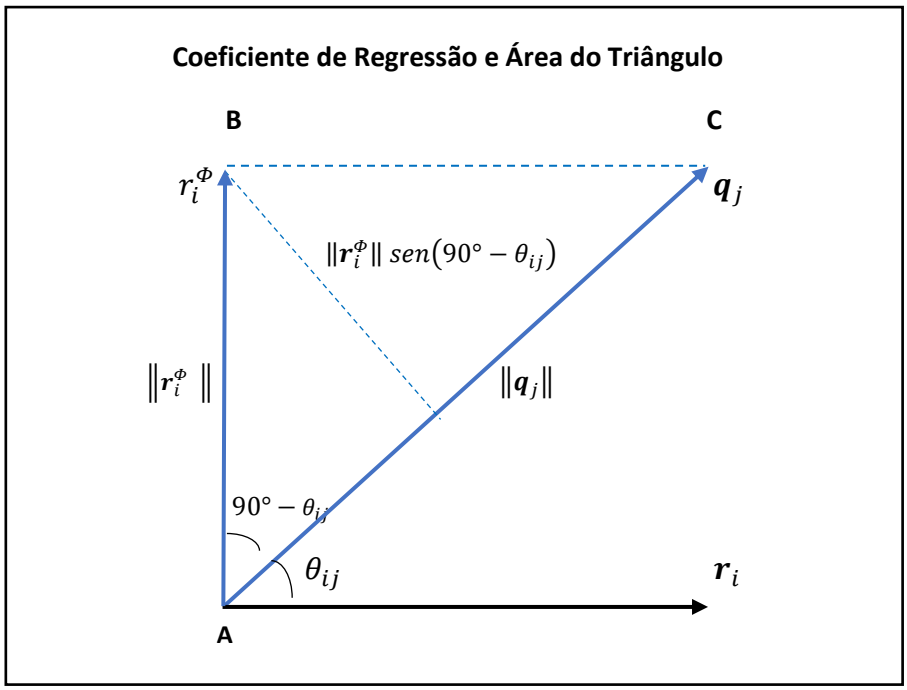

Figura 3.9: O coeficiente de regressão $b_{ij} = \boldsymbol{r}_i' \boldsymbol{q}_j = \|\boldsymbol{r}_i\| \|\boldsymbol{q}_j\| cos\theta_{ij} = \|r_i^{\Phi}\| \|\boldsymbol{q}_j\| sen(90° - \theta_{ij})$ e é igual a duas vezes a área do triângulo $A\hat{B}C$, cujos vértices são a origem e as extremidades de r_i^{Φ} e q_j.

Da Tabela 3.8, extrai-se a matriz $\boldsymbol{R}$ do PLS aplicado à base de dados *Oliveoil*. Com auxílio da matriz de rotação $\Phi_{90°}$, as linhas de $\boldsymbol{R}$ são rotacionadas $90°$ e constituirão os novos pontos *biplot* para a construção dos triângulos (um triângulo por cada linha de $\boldsymbol{R}$). Assim,

$$R^{\Phi} = R\, \Phi_{90°} = \begin{pmatrix} -0.22 & -0.76 \\ -0.54 & 0.48 \\ -0.56 & 0.26 \\ -0.50 & -0.14 \\ -0.30 & -0.34 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0.76 & -0.22 \\ -0.48 & -0.54 \\ -0.26 & -0.56 \\ 0.14 & -0.50 \\ 0.34 & -0.30 \end{pmatrix}.$$

Com a rotação $90°$ dos r_i, o *biplot* PLS dos coeficientes de regressão assume a forma visualizada na Figura 3.10, a qual foi acrescentado o reticulado e a escala dos eixos coordenados, que irão permitir a obtenção de estimativas visuais das áreas dos triângulos posteriormente:

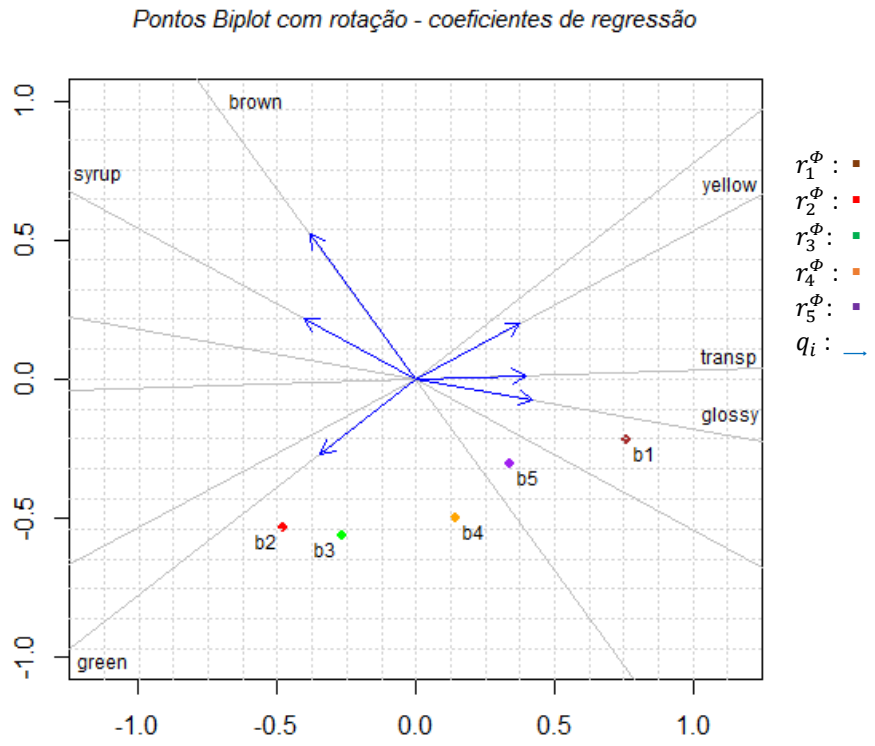

Figura 3.10: Rotação $90°$ de r_i, obtido pela multiplicação da matriz ***R*** pela matriz de rotação $\Phi_{90°}$.

Considerando a variável dependente *Brown* (terceira variável do vetor ***Y***), os coeficientes de regressão podem ser estimados visualmente a partir do *biplot* PLS que aparece na Figura 3.11. Com a finalidade de facilitar as leituras no gráfico, apenas o eixo *biplot* associado à variável de interesse é mostrado. Cada r_i' é conectado à origem, ponto $(0, 0)$, e à extremidade de $\boldsymbol{q}_3' = (-0.38 \;\; 0.52)$ para formar os triângulos que irão servir de apoio para a análise e estimação dos parâmetros. Conforme pode ser comprovado na Figura 3.11, todos os triângulos possuem uma base comum, que é $\|\boldsymbol{q}_3'\|$. Esta é uma característica geral a todos os triângulos construídos por esse método, os triângulos partilham a mesma base $\|\boldsymbol{q}_j'\|$.

As áreas dos cinco triângulos e, consequentemente, os valores dos coeficientes aproximados $\tilde{b}_{ij}$ serão proporcionais às alturas de cada triângulo. Ou seja, quanto mais afastados os pontos *biplot* rotacionados estiverem do eixo em questão, maior será a área S_{ij} e maior será o valor de $\tilde{b}_{ij}$. Daí que, como resultado, esse afastamento fornece um

indicativo da importância de alterações da variável explicativa conexa ao $\boldsymbol{r_i^{\Phi}}$ na variável resposta Y_j.

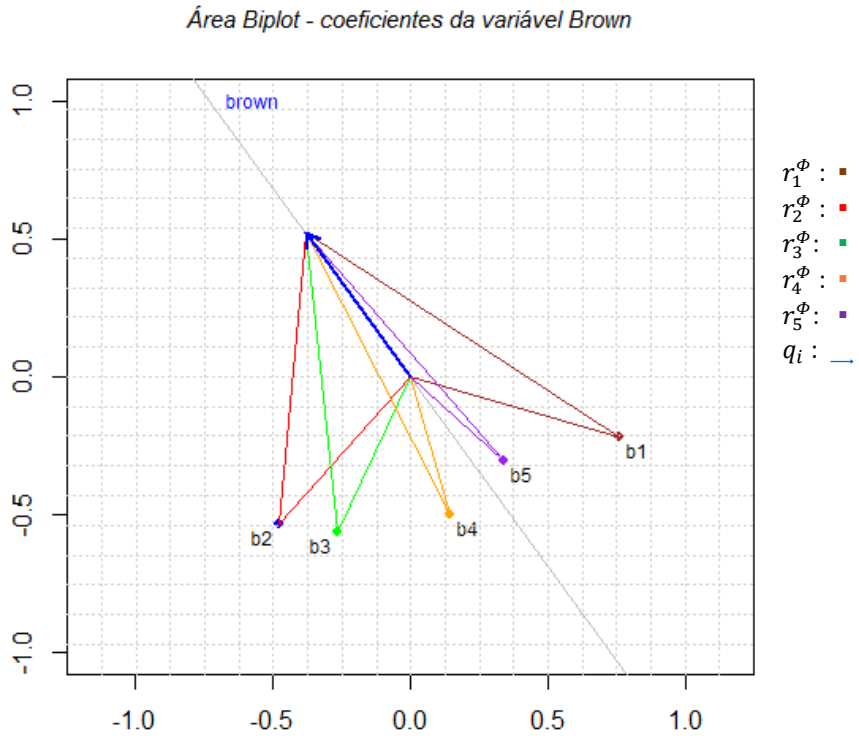

Figura 3.11: Triângulos formados após a rotação $90°$ dos r_i, em que o dobro de suas áreas fornece estimativas para os coeficientes de regressão. Pontos mais afastados do eixo *biplot* indicam que as variáveis explicativas a eles associadas têm mais importância nos efeitos sobre Y_3 (variável *Brown*).

Além disso, considerando o sentido do vetor *biplot*, o coeficiente de regressão PLS terá sinal positivo se r_i^{Φ}estiver à esquerda do eixo *biplot*. Caso contrário, terá sinal negativo. Isso ocorre porque, quando $0 < (90° - \theta_{ij}) < \pi$, tem-se[4] $sen\,(90° - \theta_{ij}) > 0$. Isto posto, verifica-se na Figura 3.11 que, como os pontos *biplot* r_2^{Φ}, r_3^{Φ} e r_4^{Φ} estão à esquerda do eixo de *Brown*, então os coeficientes $\tilde{b}_{23}$, $\tilde{b}_{33}$ e $\tilde{b}_{43}$ são positivos, enquanto que $\tilde{b}_{13}$e $\tilde{b}_{53}$ são negativos por r_1^{Φ} e r_5^{Φ}estarem posicionados à direita. Ademais, como r_2^{Φ} é o ponto *biplot* mais afastado ortogonalmente do eixo *biplot*, isso indica que a variável resposta Y_3 será mais sensível a oscilações no preditor X_2, positivamente inclusive, do que em relação a mudanças em X_4, cujo peso r_4^{Φ} está mais próximo do eixo. Em outras palavras, uma vez que a área do triângulo vermelho (b_2), na Figura 3.11, é maior do que a área do triângulo amarelo (b_4), então $\tilde{b}_{23} > \tilde{b}_{43}$.

Da mesma forma, variações na variável X_5 ocasionará uma modificação pequena na variável Y_3, dado que r_5^{Φ} é o ponto *biplot* mais próximo do eixo, além do que esses efeitos

[4] Ou $(90° + \theta_{ij})$.

serão opostos por causa do sinal negativo de $\tilde{b}_{53}$. Essas interpretações são corroboradas pelos resultados calculados pelo PLS e podem ser comparadas com os valores da Tabela 3.9 (coluna *Brown*).

	yellow	***brown***
Acidity	-0.23	-0.31
Peroxide	-0.11	0.45
K232	-0.16	0.35
K270	-0.22	0.12
DK	-0.18	-0.06

Tabela 3.9: Coeficientes de regressão PLS das variáveis dependentes *yellow* e *brown*.

Para $j = 1$, repetindo-se o procedimento para a variável resposta Y_1, nomeada como *yellow*, o *biplot* PLS com os triângulos alusivos aos seus coeficientes de regressão está representado na Figura 3.12. Fazendo uso das escalas dos eixos coordenados, $0{,}4$ parece ser uma boa estimativa para o tamanho do vetor *biplot*, base dos triângulos, enquanto $0{,}5$ pode ser considerado um valor bastante aproximado para a distância de r_4^{Φ} até o eixo *biplot yellow*. Assim:

$$\tilde{b}_{41} \cong -2\ x\ \frac{(0{,}4\ x\ 0{,}5)}{2} = -0{,}2\ ,$$

que é um valor próximo da estimativa $\hat{b}_{41}$assinalada na Tabela 3.9 ($\hat{b}_{41} \cong -0.22$).

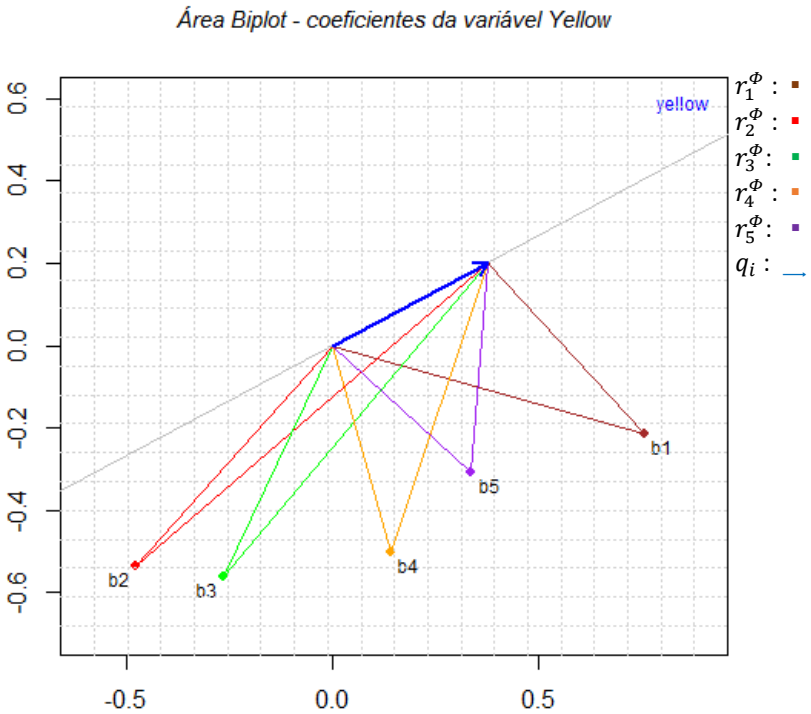

Figura 3.12: Triângulos formados após a rotação dos r_i em $90°$, em que o dobro de suas áreas fornece estimativas para os coeficientes de regressão de Y_1.

Capítulo 4

Aplicação a dados quimiométricos

4.1 Os dados de natureza química e a Estatística

Notadamente nas últimas décadas, o uso de computadores interligados a equipamentos sofisticados de análise de dados químicos produziu uma enorme quantidade de informação, com a característica marcante de possuírem um grande número de variáveis quando são geradas amostras em determinados experimentos. Esse atributo impulsionou o surgimento da Quimiometria, disciplina cujo propósito é a análise de dados químicos de natureza multivariada [7]. Embora se utilize de métodos e técnicas da Matemática e da Estatística, a Quimiometria é um assunto da Química. Assim, o objetivo desta aplicação é, puramente, mostrar o emprego de alguns dessas técnicas estatísticas, nomeadamente a regressão PLS e sua visualização *biplot*, sem ter como finalidade estabelecer discussões teóricas inerentes a outras áreas do saber. Contudo, algumas definições básicas acerca dos dados que serão trabalhados neste estudo são inicialmente colocadas.

Ao longo deste capítulo, visando a aplicação do método PLS para a extração das componentes e posterior construção de seu *biplot*, será considerada uma base de dados fornecida por Iola M. F. Duarte e Ana Maria P. C. Gil, ambas do Departamento de Química da Universidade de Aveiro. Essa base, que denominaremos *Base Plasma*, contém dados espectrais de ressonância magnética nuclear (RMN) relativos a fluidos humanos de amostras de plasma sanguíneo de 25 indivíduos. A matriz das variáveis explicativas é formada pelos dados espectrais RMN de 23 variáveis representativas dos níveis de diferentes compostos presentes nos biofluidos, conforme Tabela 4.1. A matriz das variáveis dependentes é composta por duas variáveis: a idade materna e o índice de massa corporal (BMI[1]) dos 25 indivíduos[2].

[1] Do inglês *Body Mass Index.*

[2] Originalmente, a base de dados possuía 30 observações. Contudo, como há 5 registros com dados omissos nas variáveis respostas, optou-se por considerar apenas as 25 observações com dados completos.

Com o intuito de proporcionar uma maior compreensão acerca desses dados, registre-se que a espectroscopia é uma técnica utilizada para investigar a estrutura e a dinâmica de sistemas bioquímicos, como o peso molecular e a composição atômica, em que um fóton de luz provoca uma transição do estado basal de uma molécula para um estado estimulado [19]. À resposta como função da frequência desse fóton de luz dá-se o nome de espectro, ou seja, o resultado gráfico da transição entre os dois níveis de energia. No caso da espectroscopia RMN, em vez de ocorrer a absorção de energia por um elétron, um campo magnético promove a rotação do núcleo (*spin*) do estado basal para o estado de estímulo, o que irá gerar o espectro e, posteriormente, será convertido em variável por intermédio da frequência ou outro sinal específico [21]. Essa é, muito resumidamente, a técnica que deu origem aos dados da aplicação do PLS nesta dissertação.

Variáveis Explicativas	**Variáveis de Respostas**
X_1: CH3 lipids	Y_1: Maternal Age
X_2: Valine	Y_2: BMI
X_3: (CH2)n lipids	
X_4: Alanine	
X_5: CH2CH2CO lipids	
X_6: Citrulline	
X_7: CH2C=C lipids	
X_8: N-acetyl glycoprot	
X_9: CH2CO lipids	
X_{10}: Glutamine	
X_{11}: Citrate	
X_{12}: C=CCH2C=C lipids	
X_{13}: Albumin-lysyl	
X_{14}: Creatine	
X_{15}: Creatinine2	
X_{16}: Dimethylsulfone	
X_{17}: Choline	
X_{18}: Lactate	
X_{19}: Unknown1 br	
X_{20}: HC=CH lipids	
X_{21}: Tyrosine	
X_{22}: Histidine1	
X_{23}: Glucose2	

Tabela 4.1: As variáveis explicativas representando os compostos químicos e as variáveis dependentes representando características dos indivíduos da amostra da base *Plasma.*

4.2 Preparação dos dados para a aplicação

Com a utilização do algoritmo NIPALS PLS2, o que se pretende neste capítulo é a obtenção da decomposição das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$, tal que $\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E}$ e $\boldsymbol{Y} = \boldsymbol{UQ}' + \boldsymbol{F}$, o que permitirá se predizer $\boldsymbol{Y}$ pelo modelo $\widehat{\boldsymbol{Y}} = \hat{b}_1\boldsymbol{T}_1 + \hat{b}_2\boldsymbol{T}_2 + \cdots + \hat{b}_k\boldsymbol{T}_k$ ou, no espaço de $\boldsymbol{X}$, por intermédio de $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}_{PLS}$.

No Capítulo 2, foi visto que o método PLS se inicia com uma fase de pré-processamento, quando as variáveis são centradas pela média das colunas das variáveis. Nessa fase do estudo, os dados da *Base Plasma* serão centrados e também normalizados, o que significa dizer que os elementos das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ terão subtraídas a média da coluna a qual pertencem, assim como divididos pelo desvio padrão da coluna.

A aplicação do PLS a esses dados será dividida em duas etapas. Na primeira, chamada de etapa de treino, é utilizada apenas uma parte do conjunto de dados e possui como objetivo se estimar os parâmetros do modelo PLS. Na etapa seguinte, ou etapa de teste, os coeficientes de regressão PLS estimados anteriormente e os dados das variáveis explicativas da parte restante são utilizados para a predição das variáveis resposta. Com isso, o conjunto de treino será formado por matrizes $\boldsymbol{X}_{20\text{x}23}$ e $\boldsymbol{Y}_{20\text{x}2}$, em que são considerados os primeiros 20 indivíduos da base *Plasma*. Os 5 indivíduos restantes formarão o conjunto de teste.

4.3 Quantidade de componentes pelo RMSEP

No PLS assume-se que o sistema sob investigação é influenciado apenas por algumas poucas variáveis subjacentes, mas esse número não é conhecido. Nesse sentido, o primeiro objetivo do método é a estimação dessa quantidade, tal como será feito nesta Seção. Anteriormente à aplicação da metodologia do PLS e construção de seu *biplot* relativo à *Base Plasma*, é necessário saber a quantidade k adequada de variáveis latentes a usar para se descrever o modelo. Isto é feito por meio do procedimento mostrado no Capítulo 2, em que se calcula uma medida de qualidade da predição do modelo para diversas quantidades de componentes k. O critério para se determinar a quantidade k ideal é subjetivo, mas incidirá entre os k que apresentem o melhor valor para a medida considerada. Para a *Base Plasma*, a medida utilizada foi o RMSEP, sendo que, para a validação cruzada, o conjunto de treino foi dividido em 4 grupos [23] com 5

indivíduos cada para se calcular o quadrado dos erros na predição (método 4-*folds*). Segundo [1], tipicamente, a qualidade tende primeiro a aumentar, para depois decrescer. Quando tal não acontece e o RMSEP cresce à medida que se aumenta a quantidade k de variáveis latentes, isto está a indicar que o modelo está se sobreajustando aos dados. Pela Tabela 4.2 (ou Figura 4.1), é exatamente isso o que ocorre quando se aplica o PLS2 à *Base Plasma*, com valores de RMSEP menores para k igual a 1 ou 2. Assim, optou-se por extrair apenas dois fatores latentes, o que significou uma redução significativa na dimensão das matrizes envolvidas.

K	RMSEP
1	13
2	13
3	15
4	13
5	14
6	15
7	15
8	15
9	16
10	16

Tabela 4.2: Determinação da quantidade de variáveis latentes do modelo por meio da medida da habilidade de predição RMSEP. A escolha pode recair na extração de uma a três componentes.

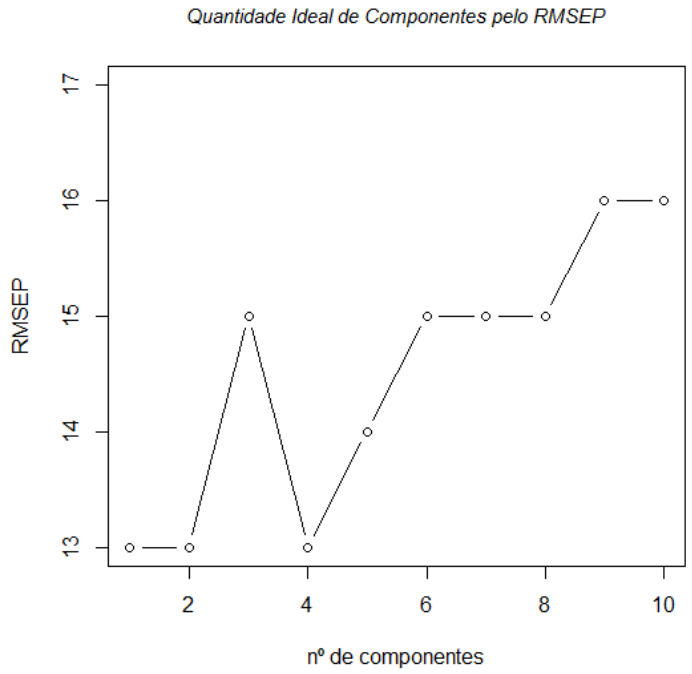

Figura 4.1: Visualização da quantidade ideal de componentes segundo o RMSEP.

4.4 Resultados e discussão do PLS

Na aplicação do PLS multivariado (PLS2) à *Base Plasma*, o que se pretende é a decomposição de $\boldsymbol{X}$ em $\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E}$ e de $\boldsymbol{Y}$ em $\boldsymbol{Y} = \boldsymbol{UQ}' + \boldsymbol{F}$, o que irá permitir a predição de $\boldsymbol{Y}$ pelo modelo $\widehat{\boldsymbol{Y}} = \hat{b}_1\boldsymbol{T}_1 + \hat{b}_2\boldsymbol{T}_2 + \cdots + \hat{b}_k\boldsymbol{T}_k$ ou, no espaço de $\boldsymbol{X}$, por intermédio de $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}_{PLS}$, em que $\widehat{\boldsymbol{B}}_{PLS} = \boldsymbol{RQ}'$, com $\boldsymbol{R} = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}$. Então, o primeiro objetivo é a extração de duas variáveis latentes para compor o modelo ($\widehat{\boldsymbol{Y}} = \hat{b}_1\boldsymbol{T}_1 + \hat{b}_2\boldsymbol{T}_2$) e, ao final, estruturar as matrizes dos *scores* $\boldsymbol{T}$ e dos *loadings* $\boldsymbol{P}$, assim como a matriz de pesos $\boldsymbol{Q}$ e a matriz de pesos $\boldsymbol{R}$. Através delas, é possível fazer a estimação dos coeficientes de regressão PLS, conceber as matrizes de aproximação[3] $\widetilde{\boldsymbol{X}}$ e $\widetilde{\boldsymbol{Y}}$, realizar as predições das respostas e, principalmente, construir o *biplots* que permitirão visualizar os parâmetros estimados do modelo ajustado aos dados pelo método PLS. As matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$, sobre as quais o PLS2 é aplicado, têm dimensões $(20\text{x}23)$ e $(20\text{x}2)$, respectivamente. Como são retidas duas componentes PLS no modelo, então as matrizes geradas pelo algoritmo têm as seguintes configurações: $\boldsymbol{T}_{20\text{x}2}$, $\boldsymbol{P}_{23\text{x}2}$, $\boldsymbol{Q}_{2\text{x}2}$ e $\boldsymbol{R}_{23\text{x}2}$. A matriz dos *scores* $\boldsymbol{T}$ está na Tabela 4.3. Esta matriz possui como característica o fato de que $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$. Esses *scores* contêm informações acerca dos indivíduos e sobre suas similaridades e dissimilaridades de acordo com o problema.

Indivíduos	*Componente 1*	*Componente 2*
1	-0.66	-1.36
2	0.52	-1.22
3	-2.05	-1.15
4	-2.32	1.32
5	-0.52	0.73
6	2.37	-1.31
7	1.10	-0.62
8	-2.08	0.64
9	-1.01	-1.22
10	1.16	1.76
11	2.50	-0.71
12	-4.41	0.82
13	5.69	0.73
14	-3.72	-1.92
15	0.90	-0.62
16	1.94	1.80
17	-1.19	-2.40
18	-0.86	2.49
19	-1.27	3.03
20	3.92	-0.78

Tabela 4.3: Variáveis latentes ortogonais T.

[3] Lembrando que $\widetilde{\boldsymbol{X}} \cong \boldsymbol{TP}'$ e que $\widetilde{\boldsymbol{Y}} \cong \boldsymbol{TQ}'$.

Os pesos em $\boldsymbol{R}$ informam sobre a articulação das variáveis na formação da relação entre X e Y, refletindo essa característica no cálculo dos coeficientes de regressão e, portanto, fornecem uma indicação da importância da variável associada. Na Tabela 4.4, onde está representada a matriz $\boldsymbol{R}$, a variável Choline (X_{17}) possui os maiores pesos nas duas componentes. Portanto, isso sugere que X_{17} poderá eventualmente ter importância diferenciada nas relações quantitativas com a idade maternal e o índice de massa corporal. A variável Citrate (X_{11}), por sua vez, teria menor importância em ambas as componentes.

Preditores	***Componente 1***	***Componente 2***
CH3 lipids	0.29	-0.01
Valine	-0.06	-0.25
X.CH2.n lipids	0.24	-0.15
Alanine	0.10	0.30
CH2CH2CO lipids	0.18	-0.16
Citrulline	-0.11	-0.18
CH2C=C lipids	0.25	-0.16
N. acetyl glycoprot	0.19	-0.22
CH2CO lipids	0.17	-0.16
Glutamine	-0.16	0.08
Citrate	-0.02	0.01
C=CCH2C=C lipids	0.39	0.23
Albumin.lysyl	-0.03	0.08
Creatine	0.27	0.39
Creatinine2	0.18	0.02
Dimethylsulfone	0.23	0.26
Choline	0.31	0.50
Lactate	0.13	0.32
Unknown1 br	0.16	0.05
HC=CH lipids	0.25	-0.08
Tyrosine	-0.20	-0.22
Histidine1	-0.29	-0.19
Glucose2	0.04	0.35

Tabela 4.4: Matriz dos pesos **R**.

A matriz de pesos $\boldsymbol{Q}$, Tabela 4.5, indica a força da relação entre as variáveis dependentes originais e as componentes. Por sua vez, a matriz $\boldsymbol{P}$ pode ser vista da mesma maneira em relação às variáveis explicativas, além de ser um bom sumário de $\boldsymbol{X}$.

Respostas	***Componente 1***	***Componente 2***
Maternal age	0.22	0.16
BMI	- 0.07	-0.37

Tabela 4.5: Matriz dos pesos **Q**.

Preditores	***Componente 1***	***Componente 2***
CH3 lipids	0.38	-0.03
Valine	0.07	-0.11
X.CH2.n lipids	0.37	-0.21
Alanine	0.04	0.37
CH2CH2CO lipids	0.32	-0.24
Citrulline	-0.12	0.14
CH2C=C lipids	0.36	-0.11
N. acetyl glycoprot	0.28	-0.13
CH2CO lipids	0.30	-0.26
Glutamine	-0.19	0.08
Citrate	-0.02	0.23
C=CCH2C=C lipids	0.34	-0.06
Albumin.lysyl	-0.07	0.30
Creatine	0.12	0.38
Creatinine2	0.18	-0.21
Dimethylsulfone	0.14	0.29
Choline	0.19	0.46
Lactate	0.05	0.39
Unknown1 br	0.21	-0.21
HC=CH lipids	0.31	-0.29
Tyrosine	-0.15	-0.02
Histidine1	-0.24	0.06
Glucose2	-0.15	0.37

Tabela 4.6: Matriz dos *loadings* **P**.

Para a predição de Y, foi preciso calcular o estimador dos coeficientes de regressão PLS que, conforme mostrado no Capítulo 2, é dado por

$$\widehat{\boldsymbol{B}}_{PLS} = \boldsymbol{R}\boldsymbol{Q}'.$$

Logo,

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}_{PLS}.$$

Dentre as possíveis interpretações do coeficiente de regressão, uma delas é que este pode ser entendido como a variação na variável de resposta correspondente a uma variação unitária do preditor, mantido o restante constante. Portanto, o coeficiente de regressão dá a importância do preditor na modelagem de Y. Na Tabela 4.7 pode ser observado que os preditores mais importantes para a modelagem da idade maternal (Y_1) são Choline (X_{17}), C=CCH2C=C lipids (X_{12}) e Creatine (X_{14}), enquanto que para o índice de massa corporal (Y_2) são Choline (X_{17}) e Creatine (X_{14}). Por sua vez, a variável Citrate (X_{11}) possui os coeficientes de menor magnitude, o que sugere ser esta a variável com menor importância para o modelo.

Preditores	***Maternal age***	***BMI***
CH3 lipids	0.06	-0.02
Valine	-0.05	0.09
X.CH2.n lipids	0.03	0.03
Alanine	0.07	-0.11
CH2CH2CO lipids	0.02	0.05
Citrulline	-0.05	0.07
CH2C=C lipids	0.03	0.04
N. acetyl glycoprot	0.01	0.07
CH2CO lipids	0.02	0.05
Glutamine	-0.02	-0.02
Citrate	-0.01	-0.01
C=CCH2C=C lipids	0.12	-0.12
Albumin.lysyl	0.01	-0.03
Creatine	0.12	-0.16
Creatinine2	0.04	-0.02
Dimethylsulfone	0.09	-0.11
Choline	0.14	-0.20
Lactate	0.08	-0.13
Unknown1 br	0.04	-0.03
HC=CH lipids	0.04	0.02
Tyrosine	-0.08	0.10
Histidine1	-0.09	0.09
Glucose2	0.06	-0.13

Tabela 4.7: Matriz das estimativas dos coeficientes B_{PLS}.

A validação do modelo foi feita com o conjunto de teste tal que $\widehat{\boldsymbol{Y}}_{teste} = \boldsymbol{X}_{teste}\widehat{\boldsymbol{B}}_{PLS}$ e o resultado é mostrado na Tabela 4.8. Verifica-se que o modelo sobrestima as respostas e apresenta um melhor desempenho em relação a variável Y_2, o índice de massa corporal.

$\boldsymbol{Y_1}$	$\widehat{\boldsymbol{Y}}_1$	$\boldsymbol{erro_1}$	$\boldsymbol{Y_2}$	$\widehat{\boldsymbol{Y}}_2$	$\boldsymbol{erro_2}$
34	41	7	21.9	24	2.1
35	37	2	21.1	24	2.9
36	38	2	25	23.9	1.1
34	37	3	22.4	23.8	1.4
28	33	5	19.8	24.2	4.4

Tabela 4.8: Validação do modelo utilizando o conjunto de teste.

4.5 Resultados e discussão do *biplot* da regressão PLS

A construção do *biplot* da regressão PLS teve como ponto de partida o cálculo dos elementos das matrizes de aproximação dos preditores e das respostas. Como as matrizes dos dados originais foram normalizados, a construção das aproximações teve que passar pelo processo inverso. Assim, foi calculada a matriz de aproximação da matriz centrada $\boldsymbol{X}$, tal que

$$\widetilde{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{P}'.$$

Em seguida, foi construída uma aproximação de cada variável explicativa original fazendo

$$dp_j\widetilde{\boldsymbol{X}}_j + \bar{x}_j\mathbf{1} \quad , \quad j = 1, \ldots, 23 \,,$$

onde dp_j é o desvio padrão da variável X_j da matriz dos dados originais, $\bar{x}_j$ é a sua média e $\mathbf{1}$ é um vetor coluna $20\text{x}1$. O mesmo procedimento foi adotado para a matriz de aproximação $\widetilde{\boldsymbol{Y}}$. Na Figura 4.2, está reproduzido o *biplot* da regressão PLS aplicado à *Base Plasma*. Os valores dos *scores* de $\boldsymbol{X}$, e que são conexos aos 20 indivíduos da amostra, estão representados por pontos *biplot* (cor laranja), enquanto que os *loadings* das 23 variáveis preditoras, que exprimem os dados espectrais, são retratados por vetores *biplot* (cor azul). Mais ainda, os pesos das variáveis dependentes foram também traçados como vetores *biplot* (cor vermelha). É possível identificar no gráfico a importância de cada variável explicativa nas componentes. Vetores como maior magnitude indicam que a variável associada possui importância em pelo menos uma das componentes, como é o caso da variável Choline (X_{17}) na componente T_2, assim como de CH3 lipids (X_1) na componente T_1. Vetores pequenos indicam pouca importância do preditor em ambas as componentes,

o que insinua que a variável poderá ter menor importância na modelagem de Y, tal qual ocorre com as variáveis Valine (X_2) e Tyrosine (X_{21}). Vetores *biplot* próximos uns aos outros indicam que as variáveis correspondentes são fortemente correlacionadas, enquanto que, quando estão a 90◦ um dos outros, denota haver fraca correlação. Assim, a variável Alanine (X_4) se mostra fortemente correlacionada à variável Lactate (X_{18}) e não correlacionada a CH3 lipids (X_1).

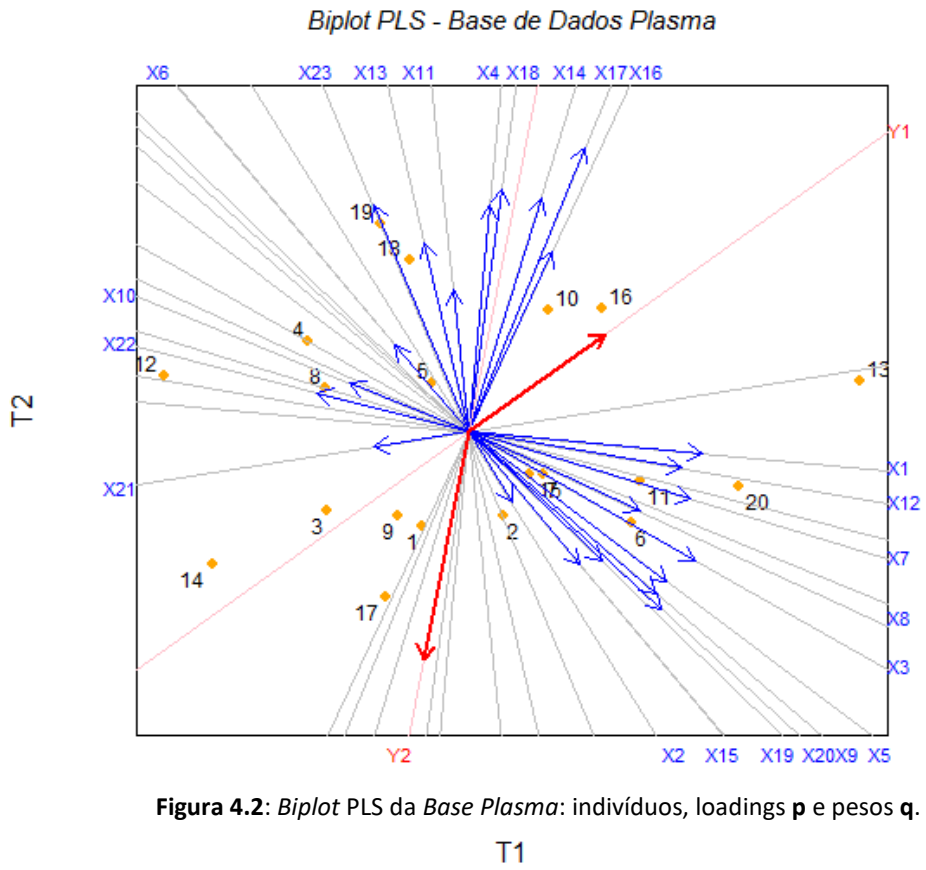

Figura 4.2: *Biplot* PLS da *Base Plasma*: indivíduos, loadings **p** e pesos **q**.

Para mostrar a calibragem dos eixos das variáveis explicativas e a leitura dos elementos da matriz de aproximação diretamente no *biplot*, foram escolhidos três preditores: Creatine (X_{14}), Creatinine2 (X_{15}) e Histidine1 (X_{22}), cujos dados espectrais estão retratados na Tabela 4.9. Seguidamente, foi concebido o *biplot* dessas variáveis e a calibração dos seus eixos *biplot* (Figura 4.3), permitindo a estimação dos elementos da matriz de aproximação a partir da projeção ortogonal dos pontos *biplot*.

Indivíduos	*Creatine*	*Creatinine2*	*Histidine1*
1	20633862	18172915	11937008
2	21106530	18604928	11479615
3	20426805	17468917	12516499
4	22642739	16084070	12883863
5	22608855	17162711	12099397
6	21555712	19447298	10726903
7	21828298	18545864	11310552
8	22080146	16539459	12714146
9	20662670	17951066	12091920
10	24048972	17356228	11532743
11	22142801	19196840	10738469
12	21573115	15450245	13665701
13	24397927	19828060	9609411
14	19238615	17143560	13106062
15	21771913	18459317	11390902
16	24314248	17669687	11223747
17	19524280	18472510	12043129
18	24145913	16116880	12417002
19	24522786	15663801	12639714
20	22490390	19843042	10160159

Tabela 4.9: Dados espectrais dos preditores cujos eixos *biplot* estão calibrados.

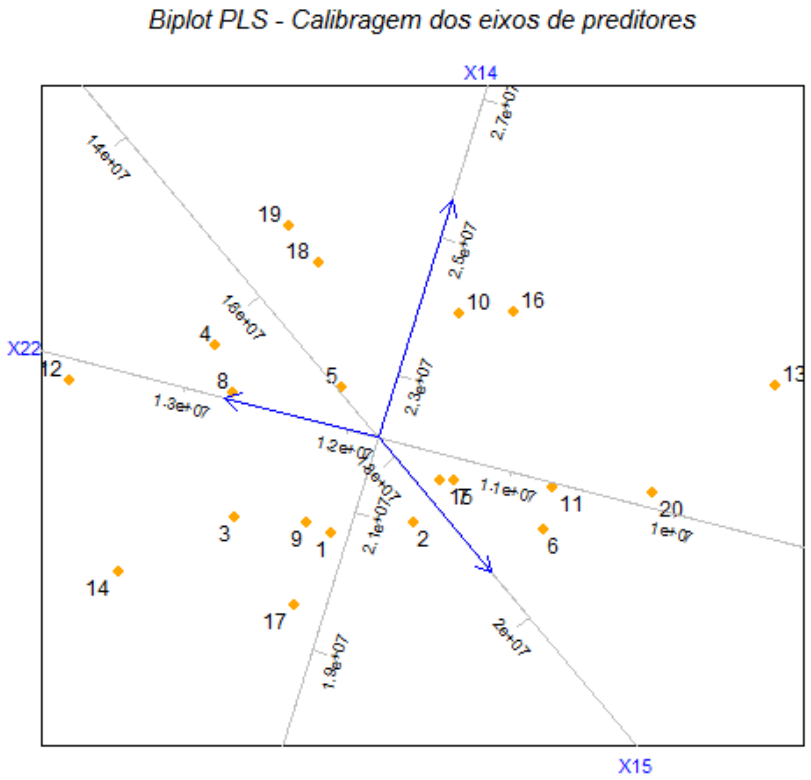

Figura 4.3: Calibragem dos eixos *biplot* das variáveis Creatine, Creatinine2 e Histidine1.

Com os eixos calibrados, os pontos *biplot* foram projetados sobre o eixo do preditor Histidine1, para que fosse possível se determinar valores aproximados daqueles calculados por $\boldsymbol{TP}'$em relação àquela variável. Assim, verifica-se na Figura 4.4, por exemplo, que a projeção do ponto referente ao indivíduo identificado com etiqueta 20 sobre o eixo *biplot* X_{22} resulta em um valor um pouco superior à marcação de 10^7. De fato, pela Tabela 4.9, o registro do espectro relativo ao indivíduo 20, no que diz respeito ao aminoácido Histidina, é 10160159.

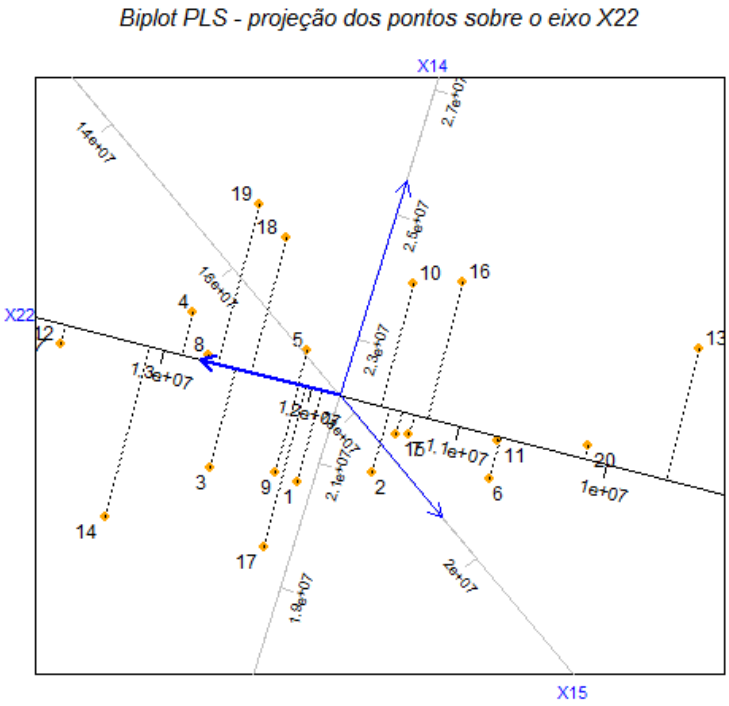

Figura 4.4: Projeção dos pontos *biplot* dos indivíduos sobre o eixo *biplot* da variável X_{22}, proporcionando a obtenção de aproximações de valores observados daquela variável explicativa.

Por seu turno, a projeção dos pontos *biplot* sobre os vetores *biplot* das variáveis dependentes fornecem uma aproximação para $\boldsymbol{Y}$. De fato, conforme mostrado no Capítulo 3, a matriz de aproximação das variáveis de respostas é tal que $\widetilde{\boldsymbol{Y}} = \boldsymbol{TQ}'$.

Para fins de comparação, a matriz $\widetilde{\boldsymbol{Y}}$ foi reproduzida na Tabela 4.10 e o *biplot* com os eixos Y_1 e Y_2 já calibrados na Figura 4.5. Na matriz $\widetilde{\boldsymbol{Y}}$, os valores calculados para a idade maternal do indivíduo 13 é 42.5 anos. Projetando-se, na Figura 4.5, o ponto *biplot* do indivíduo 13 sobre o eixo Y_1, percebe-se que o valor da leitura é bem próximo do valor 42.5.

Indivíduo	*Maternal age*	*BMI*
1	34.5	26.5
2	35.8	26
3	33.2	26.6
4	34.7	23.3
5	36.1	23.7
6	37.6	25.7
7	36.8	25.1
8	34.5	24.2
9	34.2	26.4
10	38.6	21.8
11	38.2	24.8
12	32.2	24.6
13	42.5	22
14	30.9	28.1
15	36.6	25.1
16	39.5	21.5
17	33.2	28.1
18	37.1	21.4
19	37.1	20.8
20	39.6	24.5

Tabela 4.10: As duas colunas da direita formam a matriz de aproximação $\widetilde{\boldsymbol{Y}} = \boldsymbol{TQ}'$.

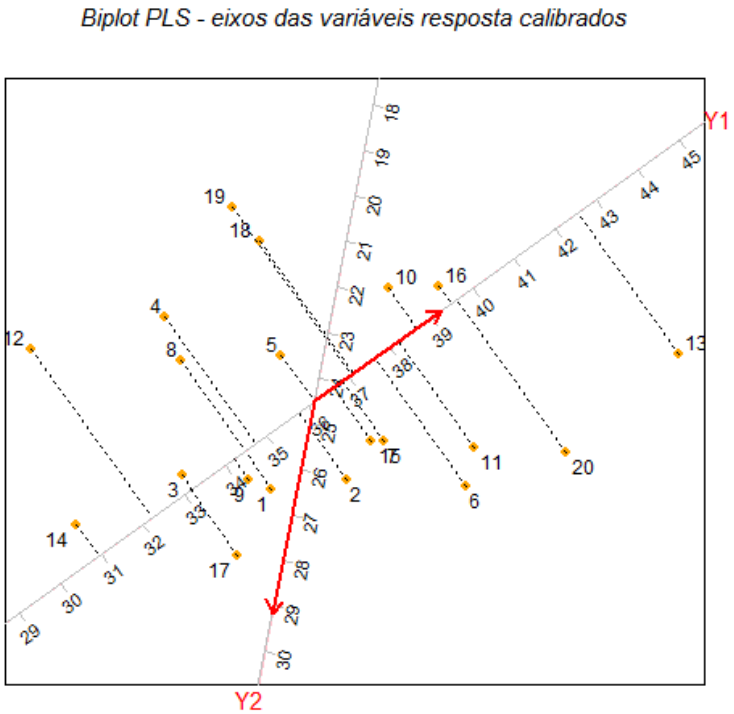

Figura 4.5: Calibragem dos eixos *biplot* das variáveis de respostas Maternal age e BMI.

Foi também construído o *biplot* que permite estimar os coeficientes de regressão do modelo PLS ajustado aos dados. Esse *biplot*, que pode ser visualizado na Figura 4.6, é construído a partir dos elementos da matriz de pesos $\boldsymbol{R}$ (Tabela 4.4) e da matriz de pesos $\boldsymbol{Q}$ (Tabela 4.5). A matriz $\boldsymbol{R}$ gerada pelo PLS2 da *Base Plasma* tem dimensão $(23\text{x}2)$ e, portanto, cada uma de suas linhas r_i , $i = 1, \dots, 23$ (23 variáveis preditoras), está associada ao coeficiente de regressão da variável preditora correspondente. Essas linhas foram marcadas no gráfico como pontos *biplot* e identificados com a etiqueta do coeficiente respectivo. Os vetores *biplot*, por outro lado, foram traçados a partir dos valores dos pesos das variáveis Maternal age e BMI, ou seja, das linhas q_i , $i = 1, 2$, da matriz $\boldsymbol{Q}$. Na Figura 4.6, o ponto *biplot* com etiqueta b_{17} refere-se ao coeficiente de regressão associado à variável explicativa Choline. As projeções de b_{17} sobre os eixos *biplot* das variáveis respostas Maternal age e BMI permitem se obter aproximações dos coeficientes estimados, que correspondem efetivamente a valores muito próximos dos valores aproximados que constam na Tabela 4.7, que no caso são 0.14 e -0.20, respectivamente.

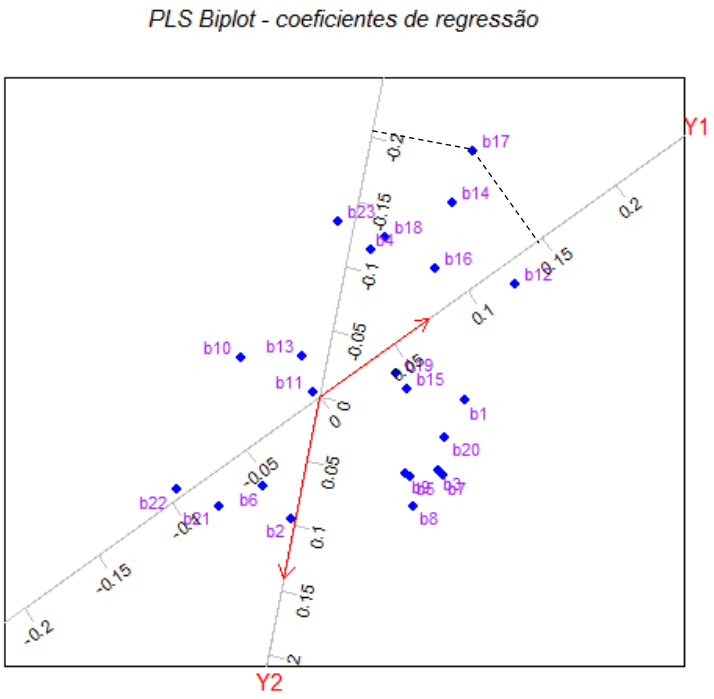

Figura 4.6: Calibragem dos eixos *biplot* das variáveis respostas Maternal age e BMI, com projeção do ponto *biplot* associado â variável preditora Choline para obtenção de uma estimativa dos coeficientes de regressão correspondente à essa variável preditora.

O *biplot* da Figura 4.6 é contundente em indicar o coeficiente de regressão b_{17}, do preditor *Choline*, como o mais influente (positivamente) em relação a Y_1, sendo que b_{22}, do preditor *Histidine1*, é o mais influente negativamente. Em relação a Y_2, o coeficiente mais influente (positivamente) é b_{21} (*Tyrosine*) e b_{17} é o mais influente negativamente.

4.6 Resultados e discussão do método área *biplot*

Foi discutido no Capítulo 3, um método visual alternativo para se obter no *biplot*, de forma aproximada, estimativas dos coeficientes de regressão resultantes do método PLS, sem que seja necessária a calibragem dos eixos *biplot* [17]. Trata-se do método que utiliza áreas de triângulos específicos traçados no gráfico. Considerando os resultados do PLS em relação aos dados espectrais RMN da *Base* Plasma, foi feita a rotação da matriz $\boldsymbol{R}$ com o auxílio da matriz de rotação

$$\boldsymbol{\Phi}_{90^\circ} = \begin{pmatrix} \cos 90^\circ & sen\ 90^\circ \\ -sen\ 90^\circ & \cos 90^\circ \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Deste modo, foi obtida uma nova matriz $\boldsymbol{R}^{\Phi} = \boldsymbol{R}\,\boldsymbol{\Phi}_{90^\circ}$, que se encontra descrita na Tabela 4.11.

Preditores	*Componente 1*	*Componente 2*
CH3 lipids	0.01	0.29
Valine	0.25	-0.06
X.CH2.n lipids	0.15	0.24
Alanine	-0.30	0.10
CH2CH2CO lipids	0.16	0.18
Citrulline	0.18	-0.11
CH2C.C lipids	0.16	0.25
N. acetyl_glycoprot	0.22	0.19
CH2CO lipids	0.16	0.17
Glutamine	-0.08	-0.16
Citrate	-0.01	-0.02
C.CCH2C.C lipids	-0.23	0.39
Albumin.lysyl	-0.08	-0.03
Creatine	-0.39	0.27
Creatinine2	-0.02	0.18
Dimethylsulfone	-0.26	0.23
Choline	-0.50	0.31
Lactate	-0.32	0.13
Unknown1 br	-0.05	0.16
HC.CH_lipids	0.08	0.25
Tyrosine	0.22	-0.20
Histidine1	0.19	-0.29
Glucose2	-0.35	0.04

Tabela 4.11: Matriz $\boldsymbol{R}^{\Phi}$ dos pesos ajustados após rotação da matriz **R** de 90◦.

As linhas dessa matriz $\boldsymbol{R}^{\Phi}$ passaram então a ser os novos pontos *biplot*, sendo que a ligação de cada um deles à origem e a um vetor *biplot* passaram a formar triângulos, conforme pode ser visto na Figura 4.7, cujo dobro das áreas fornecem estimativas para os coeficientes de regressão.

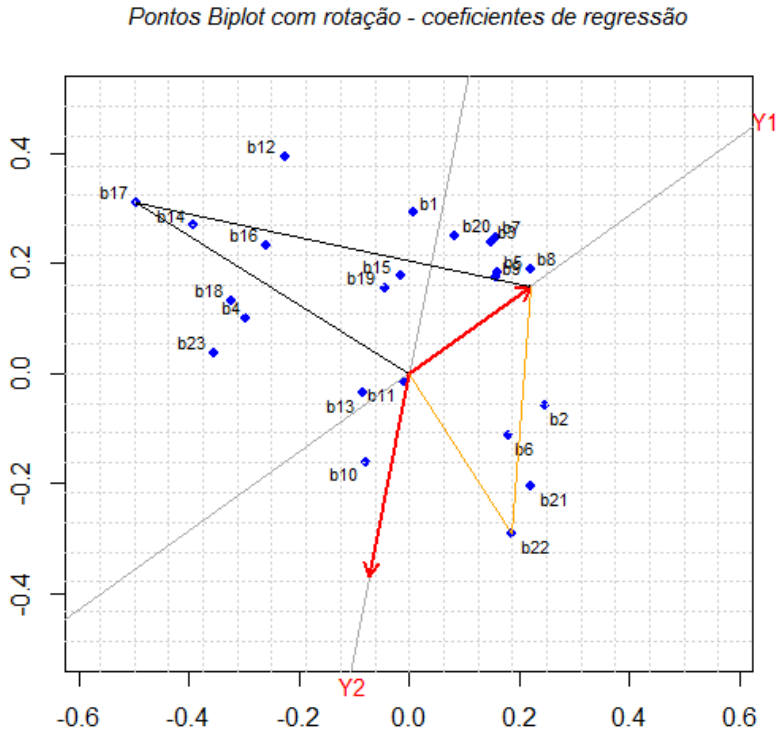

Figura 4.7: As linhas de $\boldsymbol{R}^{\Phi}$ como os novos pontos *biplot* e os triângulos formados pelos pontos, origem e vetores *biplot*. Apenas dois triângulos (com os correspondentes coeficientes de regressão) foram desenhados para não sobrecarregar o gráfico.

Especificamente, observe-se na Figura 4.7 que, após terem sofrido uma rotação de 90◦, os pontos *biplot* referentes às variáveis Choline (X_{17}) e Histedine1 (X_{22}) foram conectados à origem dos eixos das componentes e também ao vetor dos pesos da variável Maternal age (Y_1), formando dois triângulos. Esses polígonos têm em comum uma mesma base, o segmento que representa o vetor $\boldsymbol{q}_1$. Visualmente, pela Figura 4.8, é possível se verificar que a área do triângulo na cor verde é maior do que a área do triângulo cinzento e, portanto, isso indica que o coeficiente de regressão da variável Choline para Y_1 é maior, em termos absolutos, do que o coeficiente de Histidine1 para a mesma variável resposta.

As marcações nos eixos coordenados permitem, inclusive, o cálculo aproximado dessas áreas, dado que os segmentos no interior do *biplot* estão na mesma escala. Sabe-se que o

sinal dos coeficientes de regressão é dado pela posição do ponto triângulo em relação ao eixo *biplot* e ao sentido do vetor *biplot*. Triângulos à esquerda do eixo indicam que o coeficiente tem sinal positivo, sendo negativo caso contrário. Portanto, o coeficiente de Choline deve ser positivo, enquanto que o coeficiente de Histidine1 deve ser negativo. Pela Tabela 4.7 , é possível se verificar que as estimativas dos coeficientes PLS são $\hat{b}_{17} = 0.14$ e $\hat{b}_{22} = -0.09$, o que significa que essa análise está de acordo.

Outro ponto de destaque é que, como a base é comum a todos os triângulos em relação a uma mesmo variável dependente, o coeficiente será maior ou menor dependendo da distância (ortogonal) dos pontos ao eixo *biplot*, dado que essa representa a altura dos triângulos. Assim, nesse contexto, pontos *biplot* muito próximos ao eixo *biplot* indicam que os coeficientes de regressão estão muito próximo de zero, como é o caso de b_8 , b_{11} e b_{13} mostrados na Figura 4.7. De fato, as variáveis N-acetyl-glycoprot (X_8), Citrate (X_{11}) e Albumin-lysyl (X_{13}) possuem os menores coeficientes de regressão para Y_1 de acordo com a Tabela 4.7.

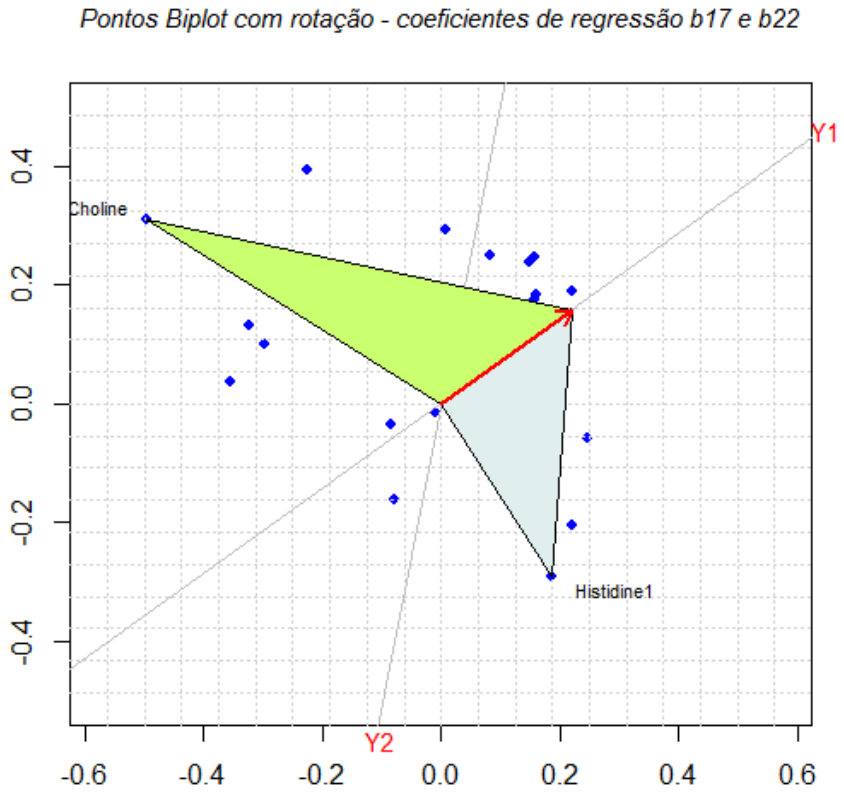

Figura 4.8: As áreas dos triângulos formados pelos pesos r^{Φ} das variáveis Choline e Histidine1 e pelos pesos *q* da variável resposta Maternal age.

Além de b_{17}, existem pontos que estão afastados de ambos os eixos *biplot*, como b_{12}, b_{14} e b_{16} (Figura 4.7), indicando que as variáveis preditoras associadas a esses pontos são influentes na predição tanto da resposta *Maternal Age* quanto de BMI. Em relação aos pontos exemplificados, essas variáveis são C=CCH2C=C lipids (X_{12}), Creatine (X_{14}) e Dimethylsulfone (X_{16}). Contudo, os efeitos exercidos por essa influência são diferentes em relação a cada uma das variáveis dependentes. Visto que b_{12}, b_{14} e b_{16} estão posicionados do lado esquerdo do eixo *biplot* de Y_1 (*Maternal Age*), eventual variação positiva[4] em alguma das variáveis associadas (por exemplo, C=CCH2C=C lipids) levará, *ceteris paribus*, a uma variação positiva em *Maternal Age*.

Já em relação a Y_2 (BMI), ocorre justamente o contrário, dado que b_{12}, b_{14} e b_{16} estão à direita do eixo respectivo (considerando o sentido do vetor *biplot*). Portanto, mantido todo o restante constante, uma variação positiva em Creatine (X_{14}) acarretará em uma variação negativa em BMI, assim como um decréscimo em Creatine leva a uma variação positiva na variável resposta Y_2.

4.7 Fechamento das discussões

Primeiramente, quanto ao estágio de construção do modelo, verifica-se a importância da representação gráfica do *biplot* PLS (Figura 4.2). A custa deste, é possível deduzir a estrutura de correlação subjacente aos dados pela caracterização de dois grupos de vetores *biplot* bem definidos, o que significa também a existência de dois grupos de variáveis explicativas fortemente correlacionadas. É possível se observar em cada um desses grupos que os vetores estão muito próximos uns dos outros. Isso indica que as variáveis associadas a esses vetores são positivamente correlacionadas. É o caso, por exemplo, das variáveis preditoras CH3 lipids (X_1) e Valine (X_2), cujo ângulo entre estes vetores está próximo de zero. Acrescente-se que, quanto mais próximo de 0° estiver o ângulo entre os vetores, mais forte será a correlação, que se dará de forma positiva. Como exemplo, verifique-se o caso de CH2CO lipids (X_9) e HC=CH lipids (X_{20}), cujos vetores estão quase sobrepostos.

O primeiro grupo de variáveis positivamente correlacionadas entre si é formado, basicamente, pelas variáveis preditoras associadas aos lipídios, cujas representações estão localizadas entre os vetores *biplot* de X_1 e X_2 (inclusive) na Figura 4.2. Note-se que são

[4] Assim como, neste caso, uma variação negativa na variável preditora levará a uma variação negativa na resposta (mantidos inalterados os restantes).

essas as variáveis com maior importância na componente T_1, conforme pode ser verificado pelos pesos constantes na Tabela 4.4.

O outro grupo importante, com variáveis correlacionadas positivamente entre si, está delimitado pelos vetores *biplot* associados às variáveis preditoras X_{23} e X_{16} (Figura 4.2). Note-se, por outro lado, que entre os dois grupos existem vetores que formam ângulos próximos a $90°$, indicando que as variáveis preditoras associadas não são correlacionas, como, por exemplo, CH2C=C lipids (X_7) e Choline (X_{17}).

Quanto aos resultados da estimação dos coeficientes de regressão PLS e considerando a etapa que envolve a predição, a Tabela 4.7 sugere que algumas variáveis exercem pouca influência na modelação das respostas. É o caso das variáveis Citrate (X_{11}) e Albumin.lysyl (X_{13}), que apresentam coeficientes próximos de zero em relação às duas variáveis dependentes. Tal conclusão também podem ser extraídas recorrendo-se à análise do *biplot* dos coeficientes de regressão (Figura 4.6) ou do gráfico área *biplot* (Figura 4.7). Dos *biplots* apresentados, fica evidenciada a importância da variável Choline (X_{17}) para o modelo.

Colocadas todas essas considerações, é possível se concluir que o *biplot* PLS é uma ferramenta exploratória útil e eficaz para a visualização dos resultados do método PLS, seja na compreensão da estrutura subjacente após a decomposição $\widetilde{\boldsymbol{X}} = \boldsymbol{TP}'$ e $\widetilde{\boldsymbol{Y}} = \boldsymbol{TQ}'$, seja na interpretação do modelo preditivo $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}_{PLS}$. Quanto aos métodos apresentados, percebe-se que a calibragem dos eixos *biplot* fornece estimativas mais precisas para os coeficientes de regressão. Contudo, quando o propósito é a interpretação, a técnica da área *biplot* se mostra mais apropriada por fornecer maior quantidade de informações.

Capítulo 5

Conclusões

Ao longo desta dissertação foi estudado o método de mínimos quadrados parciais e a construção de seus *biplots*, com sua aplicação a dados reais ao final. A abordagem do tema consistiu em, primeiramente, mostrar o funcionamento do algoritmo NIPALS para se obter os *scores* e *pesos* do PLS segundo o ponto de vista e interpretação da Estatística. O PLS é um método que projeta uma grande quantidade de pontos, a matriz $\boldsymbol{X}$, sobre um hiperplano com dimensão menor e de tal forma que as coordenadas da projeção (*scores* $\boldsymbol{t}$) são bons preditores de $\boldsymbol{Y}$, na perspectiva da regressão linear. Nesse mesmo sentido, foi mostrado que a orientação do hiperplano, expressa pelas inclinações $\boldsymbol{p}$, pode ser interpretada por meio de coeficientes de regressão, dado que os *loadings* cumprem exatamente esta função na regressão de $\boldsymbol{X}$ sobre $\boldsymbol{T}$, ou de seus resíduos $\boldsymbol{E}$ sobre $\boldsymbol{T}$, quando o algoritmo procede a deflação da matriz das variáveis explicativas em cada iteração.

Em seguida, mas antes de começar a tratar da visualização do PLS, foram discutidos os primeiros aspectos da concepção do *biplot*, em especial a relação existente entre o produto escalar e os elementos de uma matriz que tenha sido fatorada como produto de outras duas. Foi visto, portanto, que a decomposição das matrizes de dados do PLS como produto das matrizes de *scores* e de *loadings* permite se conceber um gráfico com a representação simultânea dessas últimas, a que se dá o nome de *biplot* PLS. Em complemento às funcionalidades dos *biplots* PLS como ferramenta exploratória de dados multivariados, foram vistas outras duas que ampliam suas potencialidades, como a calibragem dos eixos e o método da área *biplot*. Os exemplos analisados foram suficientes para se concluir quanto à funcionalidade de ambas, calibragem e área *biplot*.

Por último, buscou-se a síntese das metodologias estudadas com a aplicação da teoria do PLS e seus *biplots* a dados reais de natureza quimiométrica. A construção do modelo PLS revelou que os dados espectrais que formam a matriz das variáveis explicativas possuem

estrutura subjacente bem definida, com a caracterização de dois grupos de variáveis fortemente correlacionadas segundo o *biplot*. Na avaliação da capacidade preditiva do modelo, embora este não tenha mostrado precisão, revelou razoabilidade em relação aos dados de teste utilizados. A utilização do método da área *biplot* para avaliação dos coeficientes de regressão estimados pelo PLS se mostrou eficiente, exibindo corretamente os coeficientes das variáveis que mais contribuem no modelo, assim como aquelas que, por estarem próximas dos eixos das componentes, são candidatas a serem eventualmente excluídas. Sugere-se que em trabalhos futuros, o modelo seja reaplicado a novas amostras e que seja aprofundada a parte inferencial, com o objetivo de se investigar sua robustez.

Bibliografia

[1] Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics*, 2: 97-106. doi: 10.1002/wics.51.

[2] Bassani, N.; Ambrogi, F.; Coradini, D.; Biganzoli, E. (2010). Use of Biplots and Partial Least Squares Regression in Microarray Data Analysis for Assessing Association between Genes Involved in Different Biological Pathways. *Computational Intelligence Methods for Bioinformatics and Biostatistics - 7th International Meeting*, Palermo, Italy. Springer, p. 123 –134.

[3] Belsley, D. A., Kuh, E., and Welsch, R. E. (2004). *Regression Diagnostics – Identifying Influential Data and Sources of Collinearity*. Wiley, Hoboken, New Jersey.

[4] Bjorn, M.; Wehrens, R.; Liland, K.H. (2015). pls: Partial Least Squares and Principal Component Regression. R package version 2.5-0. http://CRAN.R-project.org/package=pls

[5] Chatterjee, S.; Hadi, A.S. (2012). *Regression Analysis by Example.* 5.ed. Willey.

[6] Esbensen, K.H.; Guyot, D.; Westad, F.; Houmoller, L.P. (2004). *Multivariate Data Analysis – In Practice: An Introduction to Multivariate Data Analysis and Experimental Design*. 5. Ed. Camo.

[7] Ferreira, M.M.C.; Antunes, A.M.; Melgo, M.S.; Volpe, P.L.O. (1999). Quimiometria I: Calibração multivariada, um tutorial. *Química Nova*, 22(5), 724-731. https://dx.doi.org/10.1590/S0100-40421999000500016

[8] Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis*. Biometrika,* 58, p.453-467*.*

[9] Garthwaite, P.H. (1994). An interpretation of Partial Least Squares. *Journal of the American Statistical Association*, March 1994, Vol. 89, No. 425.

[10] Geladi, P.; Kowalsky, B.R. (1986). Partial Least Squares Regression: A Tutorial. *Analytica Chimica Acta*, 186 (1986) 1-17.

[11] Greenacre, M. (2010). *Biplots in Practice*. FBBVA.

[12] Hoeskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, v.2, p. 211–228.

[13] Kalivas, J.H. (1997). Two Data Sets of Near Infrared Spectra. *Chemometrics and Intelligent Laboratory Systems*, v.37, p. 255–259.

[14] Koch, I. (2014) *Analysis of Multivariate and High-Dimensional Data*. Cambridge.

[15] Kuhn, M.; Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[16] Martens, H.; Naes, T. (1989). *Multivariate Calibration*. London: Wiley.

[17] Oyedele, O.F. (2014). *The Construction of a Partial Least Squares Biplot.* PhD thesis, University of Cape Town.

[18] R Core Team (2015). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[19] Rule, G.S.; Hitchens, T.K. (2006). *Fundamentals of Protein NMR Spectroscopy*, Springer.

[20] Sena, M.M.; Poppi, R.J. (2000). Avaliação do uso de métodos quimiométricos em análises de solos – *Química Nova*, 23(4).

[21] Wold, S.; Trygg, J.; Berglund, A.; Antti, H. (2001). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, v. 58, p. 131–150.

[22] Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiaah (Ed.). *Multivariate Analysis*. (pp.391-420) New York: Academic Press.

[23] Wold, S.; L. Eriksson; J. Trygg; N. Kettaneh. (2004). The PLS method - partial least squares projections to latent structures - and its applications in industrial RDP (research, development, and production). Umea, Sweden: Umea University.

[24] Yan, W.; Kang, M.S. (2003). *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists.* CRC Press.

Apêndice

Este apêndice reúne os scripts utilizados na análise dos dados que fazem parte desta dissertação.

Capítulo 2

Script do gráfico da página 23.

```
#package
library(ChemometricsWithR)
#dados
data(gasoline, package="pls")
wavelengths<-seq(900, 1700,by=2)
matplot(wavelengths,t(gasoline$NIR),lty=1,xlab="comprimento da onda(nm)",ylab="log(1/R)")
title("Espectro de absorção NIR (gasoline)",cex.main = 0.9,   font.main= 3)
```

Script dos gráficos das página 25-26.

```
#  PLS1 -NIPALS e Validação Cruzada (deixando um fora)
#dados
data(gasoline, package="pls")
# Extraindo componentes de 1 até K
K = 10
PRESS = numeric(10)
d = dim(gasoline)[1]
y.hat=matrix(0,60,K)
for (k in 1:K) {
        SQE = numeric(d)
        # Deixando uma observação (linha) fora de 1 até d = 60
        for ( l in 1:d) {

                # Centrando as matrizes de dados
                XTrain<-gasoline[-l,-1]
```

```
XTest<-gasoline[l,-1]
        n=dim(XTrain)[1]
        m=dim(XTrain)[2]
        q = dim(XTest)[1]
        x.bar = as.vector(colMeans(XTrain))
        for (i in 1:m) {
                XTrain[,i] = XTrain[,i] - x.bar[i]
                XTest[,i] = XTest[,i] - x.bar[i]
                }
yTrain<-gasoline[-l,1]
yTest<-gasoline[l,1]
        yTrain=as.matrix(yTrain)
        yTest=as.matrix(yTest)
        y.bar = sum(yTrain)/n
        yTrain = yTrain - y.bar
        # Inicializando as matrizes de scores, loadings, pesos e coeficientes
        W = matrix(0, m, k)
        T = matrix(0, n, k)
        P = matrix(0, m, k)
        b = as.matrix(numeric(k))
        # Extraindo as componentes
        for (i in 1:k) {
                W[,i] = t(XTrain) %*% yTrain
                W[,i] = W[,i] / sqrt(sum((W[,i])^2))
                T[,i] = XTrain %*% W[,i]
                vart = as.numeric(t(T[,i]) %*% T[,i])
                P[,i] = t(XTrain) %*% T[,i] / vart
                b[i] =  t(T[,i]) %*% yTrain  / vart
                XTrain = XTrain - T[,i] %*% t(P[,i])
                yTrain = yTrain - T[,i] * b[i]
                }
        t = matrix(0, q, k)
        # Deflacionando X
        for ( j in 1:i ) {
```

```
                    t[,j] = XTest %*% W[,j]
                    XTest = XTest - t[,j] %*% t(P[,j])
                    }
                # Predizendo Y
                y.hat[l,k] = t %*% b + y.bar
                # Soma dos quadrados do erros
                SQE[l] = SQE[l] + ( yTest - y.hat[l,k] )^2
            }
        # Calculando o PRESS
        PRESS[k] = sum(SQE)
}
# RMSEP e tabulação
PRESS = round(PRESS,3)
RMSEP = numeric(10)
for (i in 1:10) {
        RMSEP[i]=sqrt(PRESS[i]/d)
        }
RMSEP = round(RMSEP,3)
n.comp = rep(1:10)
res2 = cbind(n.comp, PRESS, RMSEP)
colnames(res2) = c("K", "PRESS", "RMSEP")
rownames(res2) = n.comp
res2
# Gráfico do PRESS
plot(PRESS, type="b",  xlab="nº de componentes")
title("Quantidade Ideal de Componentes - PRESS",cex.main = 0.9,   font.main= 3)
points(x=3,y=PRESS[3], cex = 1.5, pch=19, col = "blue")
# Gráfico do RMSEP
plot(RMSEP, type="b",xlim=c(1,11), ylim=c(1,8),  xlab="nº de componentes")
title("Quantidade Ideal de Componentes - RMSEP",cex.main = 0.9,   font.main= 3)
points(x=3,y=RMSEP[3], cex = 1.5, pch=19, col = "red")
# Gráficos da precisão da predição de Y para K = 1 até 4
par(mfrow = c(2, 2))
observado=as.matrix(gasoline[,1])
```

```
for (k in 1:4){
        plot(observado[,1],y.hat[,k],main = paste("Qualidade da predição quando K = ", k),
xlab="valores observados", ylab="valores preditos")
        abline(a=0,b=1,col="blue")
        }
#fim do script
```

Capítulo 3

Script do PLS2, biplot, calibragem e área biplot para os dados do ficheiro *oliveoil*.

```
# dados
data(oliveoil, package="pls")
X = oliveoil[,1]
Y = oliveoil[,2]
# Centrando os dados
Xc <- scale(X,center=TRUE,scale=TRUE)
XTrain= as.matrix(Xc)
Yc <- scale(Y,center=TRUE,scale=TRUE)
YTrain= as.matrix(Yc)
        n=dim(XTrain)[1]
        m=dim(XTrain)[2]
        #q = dim(XTest)[1]
        l=dim(YTrain)[2]
        # Quantidade de componentes pretendidas (K)
        k = 2
        # Inicializando matrizes
        W = matrix(0, m, k)
        T = matrix(0, n, k)
        C = matrix(0, l, k)
        Q = matrix(0, l, k)
        U = matrix(0, n, k)
```

```
P = matrix(0, m, k)
b = numeric(k)
        epslon = 10^-6
        for (i in 1:k) {
                u = as.matrix(YTrain[,1])
                delta.u=1
                while ( delta.u > epslon) {
                        varu= as.numeric(t(u) %*% u)
                        W[,i] = (t(XTrain) %*% u) / varu
                        W[,i] = W[,i] / sqrt(sum((W[,i])^2))
                        T[,i] = XTrain %*% W[,i]
                        vart = as.numeric(t(T[,i]) %*% T[,i])
                        C[,i] = t(YTrain) %*%  T[,i] / vart
                        U[,i] = YTrain %*% C[,i]
                        delta.u = t(U[,i]-u) %*% (U[,i]-u)
                        u = U[,i]
                        }
        b[i]=t(U[,i])%*%T[,i] / vart
        P[,i] = t(XTrain) %*% T[,i] / vart
        Q[,i] = t(YTrain) %*% U[,i] / varu
        XTrain = XTrain - T[,i] %*% t(P[,i])
        YTrain = YTrain - b[i]*T[,i] %*% t(C[,i])
        }
r = solve(t(P)%*% W)
R = W %*% r
B = R %*% t(C)
B
```

```
# Aproximação de X
        x.bar=as.vector(colMeans(X))
        sx=numeric(m)
        for (i in 1:m){
                sx[i]=sd(X[,i])
                }
```

```
X.til=(T%*%t(P))
for (i in 1:5) {
        X.til[,i]=X.til[,i]*sx[i]
        X.til[,i]=X.til[,i]+x.bar[i]
        }
X.aprox=X.til
X.aprox[,1:3]=round(X.til[,1:3],2)
X.aprox[,4:5]=round(X.til[,4:5],3)
X.aprox
```

```
# Aproximação de Y
y.bar=as.vector(colMeans(Y))
sy=numeric(l)
for (i in 1:l){
        sy[i]=sd(Y[,i])
        }
Y.til=(T%*%t(C))
for (j in 1:6) {
        Y.til[,j]=Y.til[,j]*sy[j]
        Y.til[,j]=Y.til[,j]+y.bar[j]
        }
Y.aprox
```

```
# Biplot
plot(T[,1],T[,2],pch=19,cex=0.8,col="brown", xlab=" ",ylab=" ",xaxt="n", yaxt="n",
+  xlim=c(-3,3),ylim=c(-3,3),asp=1)
textxy(T[,1],T[,2],rownames(X0),cex=0.75)
for (j in 1:6){
        abline(0, (P[j,2])/(P[j,1]), col="pink")
        }
arrows(0,0,2*P[,1],2*P[,2],length=0.1, lwd=1.5, col="red")
textxy(0.005*P[1,1],3.6*P[1,2],colnames(X0)[1],cex=0.75,col="red")
for (i in 1:6){
        abline(0, (C[i,2])/(C[i,1]), col="gray")
```

```
                }
arrows(0,0,2*C[,1],2*C[,2],length=0.1,lwd=1.5, col="blue")
        textxy(7.8*C[1,1],10*C[1,2],colnames(Y0)[1],cex=0.75, col="blue")
title("Biplot PLS ",cex.main = 0.9,   font.main= 3)

# Calibragem dos preditores
n=dim(Xc)[1]
m=dim(Xc)[2]
e = as.matrix(rep(0,m))
phi = matrix(0,n,m)
mu.cent = matrix(0,n,m)
mu = matrix(0,n,m)
dv = as.matrix(rep(0,m))
for (j in 1:m) {
        dv[j,] = sd(X[,j])
        e[j,] = 1
        den = t(e)%*%P%*%t(P)%*%e
        for (i in 1:n) {
                mu.cent[i,j] = T[i,]%*%t(P)%*%e
                phi[i,j] = mu.cent[i,j]/sqrt(den)
                }
        e[j]=0
        }
for (k in 1:5){
        mu[,k]=(mu.cent[,k]*dv[k])+x.bar[k]
        }
mu[,1:2]=round(mu[,1:2],2)
mu[,3:5]=round(mu[,3:5],3)
mu

# packages
library(MASS)
library(calibrate)
```

```
plot(T[,1],T[,2],pch=19,cex=0.8,col="brown", xlab=" ",ylab=" ",xaxt="n", yaxt="n",
+ xlim=c(-3,3),ylim=c(-3,3),asp=1)
textxy(T[,1],T[,2],rownames(X0),cex=0.75, col="purple")
        for (j in 1:6){
                abline(0, (P[j,2])/(P[j,1]), col="gray")
                }
arrows(0,0,2*P[,1],2*P[,2],length=0.1, lwd=1.5, col="red")
        textxy(0.005*P[1,1],3.6*P[1,2],colnames(X0)[1],cex=0.75,col="red")
        textxy(5.2*P[2,1],7.3*P[2,2],colnames(X0)[2],cex=0.75,col="red")
title("Biplot PLS - calibragem preditores",cex.main = 0.9,   font.main= 3)



range(mu[,1])
ticklab <- seq(-1,1,by=0.1)
ticklabc <- ticklab-mean(mu[,1])
xc <- (mu[,1]-mean(mu[,1]))
g <- P[1,1:2]
Calibrate.X1 <-
+calibrate(g,xc,ticklabc,T[,1:2],ticklab,tl=0.05,dp=T,cex.axislab=0.6,where=1,labpos=4,axiscol="gray
")

# Calibragem das respostas
n=dim(Yc)[1]
p=dim(Yc)[2]
e = as.matrix(rep(0,p))
phiy = as.matrix(rep(0,n))
muy.cent = matrix(0,n,p)
muy = matrix(0,n,p)
dvy = as.matrix(rep(0,p))
for (j in 1:p) {
        dvy[j,] = sd(Y[,j])
        e[j,] = 1
        den = t(e)%*%C%*%t(C)%*%e
        for (i in 1:n) {
                muy.cent[i,j] = T[i,]%*%t(C)%*%e
```

```
                phiy[i] = muy.cent[i,p]/den
                }
        e[j]=0
        }
for (k in 1:6){
        muy[,k]=(muy.cent[,k]*dvy[k])+y.bar[k]
        }
muy[,1:2]=round(muy[,1:2],2)
muy[,3:5]=round(muy[,3:5],3)
muy

plot(T[,1],T[,2],pch=19,cex=0.8,col="brown", xlab=" ",ylab=" ",xaxt="n", yaxt="n",
+ xlim=c(-3,3),ylim=c(-3,3),asp=1)
textxy(T[,1],T[,2],rownames(X0),cex=0.75,col="purple")
for (i in 1:6){
                abline(0, (C[i,2])/(C[i,1]), col="gray")
                }
arrows(0,0,2*C[,1],2*C[,2],length=0.1,lwd=1.5, col="blue")
        textxy(7.8*C[1,1],10*C[1,2],colnames(Y0)[1],cex=0.75, col="blue")
        textxy(8.7*C[2,1],8*C[2,2],colnames(Y0)[2],cex=0.75, col="blue")
title("Biplot PLS - eixos das variáveis resposta calibrados", cex.main = 0.9,   font.main= 3)

range(muy[,1])
ticklab <- seq(-10,120,by=10)
ticklabc <- ticklab-mean(muy[,1])
yc <- (muy[,1]-mean(muy[,1]))
g <- C[1,1:2]
Calibrate.Y1 <-
+calibrate(g,yc,ticklabc,T[,1:2],ticklab,tl=0.15,dp=F,cex.axislab=0.7,where=1,labpos=4,axiscol="gray
")

range(muy[,2])
ticklab <- seq(-20,80,by=10)
ticklabc <- ticklab-mean(muy[,2])
yc <- (muy[,2]-mean(muy[,2]))
```

```
g <- C[2,1:2]

Calibrate.Y2 <-
+calibrate(g,yc,ticklabc,T[,1:2],ticklab,tl=0.15,dp=F,cex.axislab=0.7,where=1,labpos=4,axiscol="gray
")



# Calibragem para os Coeficientes de Regressão

b.lab=c("b1","b2","b3","b4","b5")

plot(R[,1],R[,2],pch=20,cex=1.2,col="brown", xlab="",ylab="",xaxt="n", yaxt="n", xlim=c(-
1,1),ylim=c(-1,1),asp=1)

textxy(R[,1],R[,2],b.lab,cex=0.75,col="purple")

for (j in 1:6){

        abline(0, (C[j,2])/(C[j,1]), col="gray")

        }

arrows(0,0,C[,1],C[,2],length=0.1,lwd=1.5, col="blue")

        textxy(2.6*C[1,1],3.3*C[1,2],colnames(Y)[1],cex=0.75, col="blue")

        textxy(2.85*C[2,1],3.6*C[2,2],colnames(Y)[2],cex=0.75, col="blue")

title("PLS Biplot - coeficientes de regressão: Yellow", cex.main = 0.9,   font.main= 3)


range(B[,1])

ticklab <- seq(-0.8,0.8,by=0.1)

bc <- B[,1]

g <- C[1,1:2]

Calibrate.B1 <-
+calibrate(g,bc,ticklab,R[,1:2],ticklab,tl=0.05,dp=T,cex.axislab=0.6,where=1,labpos=4,axiscol="gray
")


range(B[,2])

ticklab <- seq(-0.8,0.8,by=0.1)

bc <- B[,2]

g <- C[2,1:2]

Calibrate.B2 <-
+calibrate(g,bc,ticklab,R[,1:2],ticklab,tl=0.05,dp=F,cex.axislab=0.6,where=1,labpos=4,axiscol="gray"
)


# Área biplot

M = c(0 , -1, 1, 0)
```

```
M = matrix(M,2)
M
rot = R%*%M
rot
library(MASS)
library(calibrate)
b.lab=c("b1","b2","b3","b4","b5")
cor=c("brown","red","green","orange","purple")
plot(rot[,1],rot[,2],pch=19,cex=0.8,col=cor, xlab="",ylab="", xlim=c(-1,1),ylim=c(-1,1),asp=1)
grid(nx = 20, ny = 20, col = "lightgray", lty = "dotted", lwd = par("lwd"))
textxy(rot[,1],rot[,2],b.lab,cex=0.75,col="black")
for (j in 1:6){
        abline(0, (C[j,2])/(C[j,1]), col="gray")
        }
arrows(0,0,C[3,1],C[3,2],length=0.1,lwd=1, col="blue")
textxy(2.6*C[1,1],3.3*C[1,2],colnames(Y)[1],cex=0.75, col="black")
textxy(2.85*C[2,1],3.6*C[2,2],colnames(Y)[2],cex=0.75, col="black")
textxy(1.1*C[3,1],1.84*C[3,2],colnames(Y)[3],cex=0.75, col="black")
textxy(2.3*C[4,1],1*C[4,2],colnames(Y)[4],cex=0.75, col="black")
textxy(2.45*C[5,1],3*C[5,2],colnames(Y)[5],cex=0.75, col="black")
textxy(2.5*C[6,1],3.2*C[6,2],colnames(Y)[6],cex=0.75, col="black")
title("Pontos Biplot com rotação - coeficientes de regressão", cex.main = 0.9,   font.main= 3)
arrows(0,0,x1=rot[1,1],y1=rot[1,2],length=0.01,lwd=1.5, col="brown")
arrows(rot[1,1],rot[1,2],C[3,1],C[3,2],length=0.01,lwd=1.5, col="brown")

arrows(0,0,x1=rot[2,1],y1=rot[2,2],length=0.01,lwd=1.5, col="blue")
arrows(rot[2,1],rot[2,2],C[3,1],C[3,2],length=0.01,lwd=1.5, col="blue")

arrows(0,0,x1=rot[3,1],y1=rot[3,2],length=0.01,lwd=1.5, col="green")
arrows(rot[3,1],rot[3,2],C[3,1],C[3,2],length=0.01,lwd=1.5, col="green")

arrows(0,0,x1=rot[4,1],y1=rot[4,2],length=0.01,lwd=1.5, col="orange")
arrows(rot[4,1],rot[4,2],C[3,1],C[3,2],length=0.01,lwd=1.5, col="orange")
```

```
arrows(0,0,x1=rot[5,1],y1=rot[5,2],length=0.01,lwd=1.5, col="purple")
arrows(rot[5,1],rot[5,2],C[3,1],C[3,2],length=0.01,lwd=1.5, col="purple")
#fim do script
```

Capítulo 4

Script do PLS2, biplot, calibragem e área biplot para os dados da base plasma.

```
#dados
X <-  dados[  , 4:26]
Y <-  dados[  , 2:3  ]
Xc <- scale( X, center=TRUE, scale=TRUE)
XTrain = as.matrix(Xc)
Yc <- scale( Y, center=TRUE, scale=TRUE)
YTrain = as.matrix(Yc)
n=dim(XTrain)[1]
m=dim(XTrain)[2]
l=dim(YTrain)[2]
# Quantidade de componentes (K)
        k = 2
        W = matrix(0, m, k)
        T = matrix(0, n, k)
        C = matrix(0, l, k)
        Q = matrix(0, l, k)
        U = matrix(0, n, k)
        P = matrix(0, m, k)
        b = numeric(k)
        epslon = 10^-6
        for (i in 1:k) {
                u = as.matrix(YTrain[,1])
                delta.u=1
                        while ( delta.u > epslon) {
```

```
                varu= as.numeric(t(u) %*% u)
                W[,i] = (t(XTrain) %*% u) / varu
                W[,i] = W[,i] / sqrt(sum((W[,i])^2))
                T[,i] = XTrain %*% W[,i]
                vart = as.numeric(t(T[,i]) %*% T[,i])
                C[,i] = t(YTrain) %*%  T[,i] / vart
                U[,i] = YTrain %*% C[,i]
                delta.u = t(U[,i]-u) %*% (U[,i]-u)
                u = U[,i]
                }
        b[i]=t(U[,i])%*%T[,i] / vart
        P[,i] = t(XTrain) %*% T[,i] / vart
        Q[,i] = t(YTrain) %*% U[,i] / varu
        XTrain = XTrain - T[,i] %*% t(P[,i])
        YTrain = YTrain - b[i]*T[,i] %*% t(C[,i])
        }
        r = solve(t(P)%*% W)
        R = W %*% r
        B = R %*% t(C)
        B
```

 # Biplot

```
plot(T[,1],T[,2],pch=19,cex=0.8,col="orange",xlab="T1",ylab="T2" ,xaxt="n", yaxt="n",asp=1)
textxy(T[,1],T[,2],rownames(X0),cex=0.75)
for (j in 1:23){
                abline(0, (P[j,2])/(P[j,1]), col="gray")
                }
arrows(0,0,9*P[,1],9*P[,2],length=0.1, lwd=1.5, col="blue")
        mtext("X1", side = 4, line = 0, at=-0.5, col="blue",cex=0.7,las=2)
        mtext("X2", side = 1, line = 0, at=3, col="blue",cex=0.7)
        mtext("X3", side = 4, line = 0, at=-3.4, col="blue",cex=0.7,las=2)
        mtext("X4", side = 3, line = 0, at=0.3, col="blue",cex=0.7)
        mtext("X5", side = 1, line = 0, at=6, col="blue",cex=0.7)
        mtext("X6", side = 3, line = 0, at=-4.5, col="blue",cex=0.7)
```

```
        mtext("X7", side = 4, line = 0, at=-1.8, col="blue",cex=0.7,las=2)
        mtext("X8", side = 4, line = 0, at=-2.7, col="blue",cex=0.7,las=2)
        mtext("X9", side = 1, line = 0, at=5.5, col="blue",cex=0.7)
        mtext("X10", side = 2, line = 0, at=2, col="blue",cex=0.7,las=2)
        mtext("X11", side = 3, line = 0, at=-0.7, col="blue",cex=0.7)
        mtext("X12", side = 4, line = 0, at=-1, col="blue",cex=0.7,las=2)
        mtext("X13", side = 3, line = 0, at=-1.4, col="blue",cex=0.7)
        mtext("X14", side = 3, line = 0, at=1.5, col="blue",cex=0.7)
        mtext("X15", side = 1, line = 0, at=3.7, col="blue",cex=0.7)
        mtext("X16", side = 3, line = 0, at=2.6, col="blue",cex=0.7)
        mtext("X17", side = 3, line = 0, at=2.1, col="blue",cex=0.7)
        mtext("X18", side = 3, line = 0, at=0.8, col="blue",cex=0.7)
        mtext("X19", side = 1, line = 0, at=4.5, col="blue",cex=0.7)
        mtext("X20", side = 1, line = 0, at=5.1, col="blue",cex=0.7)
        mtext("X21", side = 2, line = 0, at=-0.8, col="blue",cex=0.7,las=2)
        mtext("X22", side = 2, line = 0, at=1.3, col="blue",cex=0.7,las=2)
        mtext("X23", side = 3, line = 0, at=-2.2, col="blue",cex=0.7)
for (i in 1:2){
        abline(0, (C[i,2])/(C[i,1]), col="pink")
}
arrows(0,0,9*C[,1],9*C[,2],length=0.1,lwd=2, col="red")
        mtext("Y1", side = 4, line = 0, at=4.4, col="red",cex=0.7,las=2)
        mtext("Y2", side = 1, line = 0, at=-1, col="red",cex=0.7)
title("Biplot PLS - Base de Dados Plasma",cex.main = 0.9,   font.main= 3)

# Aproximação de X
        x.bar=as.vector(colMeans(X))
        sx=numeric(23)
        for (i in 1:23){
                sx[i]=sd(X[,i])
                }
        X.til=(T%*%t(P))
        for (i in 1:23) {
                X.til[,i]=X.til[,i]*sx[i]
```

```
            X.til[,i]=X.til[,i]+x.bar[i]
            }
        X.aprox=X.til
        X.aprox

# Aproximação de Y
        y.bar=as.vector(colMeans(Y))
        sy=numeric(2)
        for (i in 1:2){
            sy[i]=sd(Y[,i])
            }
        Y.til=(T%*%t(C))
        for (i in 1:2) {
            Y.til[,i]=Y.til[,i]*sy[i]
            Y.til[,i]=Y.til[,i]+y.bar[i]
            }
        Ya=Y.til
        Ya[,1]=round(Ya[,1],1)
        Ya[,2]=round(Ya[,2],1)

# CALIBRAGEM PREDICTORS
n=dim(Xc)[1]
m=dim(Xc)[2]
e = as.matrix(rep(0,m))
phi = matrix(0,n,m)
mu.cent = matrix(0,n,m)
mu = matrix(0,n,m)
dv = as.matrix(rep(0,m))
for (j in 1:m) {
        dv[j,] = sd(X[,j])
        e[j,] = 1
        den = t(e)%*%P%*%t(P)%*%e
        for (i in 1:n) {
            mu.cent[i,j] = T[i,]%*%t(P)%*%e
```

```
                phi[i,j] = mu.cent[i,j]/sqrt(den)
                }
        e[j]=0
        }
for (k in 1:23){
        mu[,k]=(mu.cent[,k]*dv[k])+x.bar[k]
        }
Mu


plot(T[,1],T[,2],pch=19,cex=0.8,col="orange",xlab="",ylab="" ,xaxt="n", yaxt="n",asp=1)
textxy(T[,1],T[,2],rownames(X0),cex=0.75)
        abline(0, (P[14,2])/(P[14,1]), col="gray")
        abline(0, (P[15,2])/(P[15,1]), col="gray")
        abline(0, (P[22,2])/(P[22,1]), col="gray")
        arrows(0,0,9*P[14,1],9*P[14,2],length=0.1, lwd=1.9, col="blue")
        arrows(0,0,9*P[15,1],9*P[15,2],length=0.1, lwd=1.9, col="blue")
        arrows(0,0,9*P[22,1],9*P[22,2],length=0.1, lwd=2, col="blue")
                mtext("X14", side = 3, line = 0, at=1.5, col="blue",cex=0.7)
                mtext("X15", side = 1, line = 0, at=3.7, col="blue",cex=0.7)
                mtext("X22", side = 2, line = 0, at=1.3, col="blue",cex=0.7,las=2)
range(mu[,14])
ticklab <- seq(15000000,30000000,by=2000000)
ticklabc <- ticklab-mean(mu[,14])
xc <- (mu[,14]-mean(mu[,14]))
g <- P[14,1:2]
Calibrate.X14 <-
calibrate(g,xc,ticklabc,T[,1:2],ticklab,tl=0.2,dp=F,cex.axislab=0.6,where=1,labpos=4,axiscol="gray")


range(mu[,15])
ticklab <- seq(12000000,20000000,by=2000000)
ticklabc <- ticklab-mean(mu[,15])
xc <- (mu[,15]-mean(mu[,15]))
g <- P[15,1:2]
Calibrate.X15 <-
calibrate(g,xc,ticklabc,T[,1:2],ticklab,tl=0.2,dp=F,cex.axislab=0.6,where=1,labpos=4,axiscol="gray")
```

```
range(mu[,22])

ticklab <- seq(9000000,14000000,by=1000000)

ticklabc <- ticklab-mean(mu[,22])

xc <- (mu[,22]-mean(mu[,22]))

g <- P[22,1:2]

Calibrate.X22 <-
calibrate(g,xc,ticklabc,T[,1:2],ticklab,tl=0.2,dp=T,cex.axislab=0.7,where=1,labpos=4,axiscol="black")


title("Biplot PLS - projeção dos pontos sobre o eixo X22",cex.main = 0.9,   font.main= 3)


# CALIBRAGEM RESPONSES

n=dim(Yc)[1]

p=dim(Yc)[2]

e = as.matrix(rep(0,p))

phiy = as.matrix(rep(0,n))

muy.cent = matrix(0,n,p)

muy = matrix(0,n,p)

dvy = as.matrix(rep(0,p))


for (j in 1:p) {
        dvy[j,] = sd(Y[,j])
        e[j,] = 1
        den = t(e)%*%C%*%t(C)%*%e
        for (i in 1:n) {
                muy.cent[i,j] = T[i,]%*%t(C)%*%e
                phiy[i] = muy.cent[i,p]/den
                }
        e[j]=0
        }
for (k in 1:2){
        muy[,k]=(muy.cent[,k]*dvy[k])+y.bar[k]
        }
Muy
```

```
plot(T[,1],T[,2],pch=19,cex=0.8,col="orange",xlab="",ylab="" ,xaxt="n", yaxt="n",asp=1)

textxy(T[,1],T[,2],rownames(X0),cex=0.75)

for (i in 1:2){

                    abline(0, (C[i,2])/(C[i,1]), col="pink")

                    }

arrows(0,0,9*C[,1],9*C[,2],length=0.1,lwd=2, col="red")

                    mtext("Y1", side = 4, line = 0, at=4.4, col="red",cex=0.9,las=2)

                    mtext("Y2", side = 1, line = 0, at=-1, col="red",cex=0.9)

title("Biplot PLS - eixos das variáveis resposta calibrados",cex.main = 0.9,   font.main= 3)


range(muy[,1])

ticklab <- seq(43,60,by=1)

ticklabc <- ticklab-mean(muy[,1])

yc <- (muy[,1]-mean(muy[,1]))

g <- C[1,1:2]

Calibrate.Y1 <-
calibrate(g,yc,ticklabc,T[,1:2],ticklab,tl=0.15,dp=T,cex.axislab=0.7,where=1,labpos=4,axiscol="gray"
)


range(muy[,2])

ticklab <- seq(25,40,by=1)

ticklabc <- ticklab-mean(muy[,2])

yc <- (muy[,2]-mean(muy[,2]))

g <- C[2,1:2]

Calibrate.Y2 <-
calibrate(g,yc,ticklabc,T[,1:2],ticklab,tl=0.15,dp=F,cex.axislab=0.7,where=1,labpos=4,axiscol="gray"
)

arrows(0,0,9*C[,1],9*C[,2],length=0.1,lwd=2, col="red")


# Calibragem para os coeficientes

b.lab=c("b1","b2","b3","b4","b5","b6","b7","b8","b9","b10","b11","b12","b13","b14","b15","b16","b17",
"b18","b19","b20","b21","b22","b23")

plot(R[,1],R[,2],pch=20,cex=1.2,col="blue", xlab="",ylab="",xaxt="n", yaxt="n", xlim=c(-
0.2,0.3),ylim=c(-0.5,0.6),asp=1)

textxy(R[,1],R[,2],b.lab,cex=0.65,col="purple")
```

```
for (j in 1:2){
        abline(0, (C[j,2])/(C[j,1]), col="gray")
        }
arrows(0,0,C[,1],C[,2],length=0.1,lwd=1.5, col="red")
        mtext("Y1", side = 4, line = 0, at=0.55, col="red",cex=0.9,las=2)
        mtext("Y2", side = 1, line = 0, at=-0.1, col="red",cex=0.9)
title("PLS Biplot - coeficientes de regressão",cex.main = 0.9,   font.main= 3)


range(B[,1])
ticklab <- seq(-0.2,0.3,by=0.05)
bc <- B[,1]
g <- C[1,1:2]
Calibrate.B1 <-
calibrate(g,bc,ticklab,R[,1:2],ticklab,tl=0.03,dp=F,cex.axislab=0.65,where=1,labpos=4,axiscol="gray"
)


range(B[,2])
ticklab <- seq(-0.2,0.3,by=0.05)
bc <- B[,2]
g <- C[2,1:2]
Calibrate.B2 <-
calibrate(g,bc,ticklab,R[,1:2],ticklab,tl=0.03,dp=F,cex.axislab=0.65,where=1,labpos=4,axiscol="gray"
)


arrows(0,0,C[,1],C[,2],length=0.1,lwd=1.5, col="red")


# Área biplot
M = c(0 , -1, 1, 0)
M = matrix(M,2)
rot = R%*%M
library(MASS)
library(calibrate)
b.lab=c("b1","b2","b3","b4","b5","b6","b7","b8","b9","b10","b11","b12","b13","b14","b15","b16","b17",
"b18","b19","b20","b21","b22","b23")
#b.lab=c("","","","","","","","","","","","","","","","Choline","","","","","Histidine1","")
```

```
plot(rot[,1],rot[,2],pch=19,cex=0.8,col="blue", xlab="",ylab="", xlim=c(-0.25,0.25),ylim=c(-
0.5,0.5),asp=1)
grid(nx = 20, ny = 20, col = "lightgray", lty = "dotted", lwd = par("lwd"))
textxy(rot[,1],rot[,2],b.lab,cex=0.7,col="black")
for (j in 1:2){
        abline(0, (C[j,2])/(C[j,1]), col="darkgray")
        }
arrows(0,0,C[,1],C[,2],length=0.1,lwd=2, col="red")
        mtext("Y1", side = 4, line = 0, at=0.46, col="red",cex=0.9,las=2)
        mtext("Y2", side = 1, line = 0, at=-0.1, col="red",cex=0.9)
title("Pontos Biplot com rotação - coeficientes de regressão",cex.main = 0.9,   font.main= 3)
arrows(0,0,x1=rot[17,1],y1=rot[17,2],length=0.01,lwd=1.5, col="black")
arrows(rot[17,1],rot[17,2],C[1,1],C[1,2],length=0.01,lwd=1.5, col="black")

arrows(0,0,x1=rot[22,1],y1=rot[22,2],length=0.01,lwd=1.5, col="orange")
arrows(rot[22,1],rot[22,2],C[1,1],C[1,2],length=0.01,lwd=1.5, col="orange")



v1=c(0,rot[17,1],C[1,1])
v2=c(0,rot[17,2], C[1,2])
polygon(x=v1,y=v2, col="darkolivegreen1")
v1=c(0,rot[22,1],C[1,1])
v2=c(0,rot[22,2], C[1,2])
polygon(x=v1,y=v2, col="azure2")
arrows(0,0,C[1,1],C[1,2],length=0.1,lwd=2, col="red")

#fim do script
```